Evaluating Text Classification Models

Natalie Parde
UIC CS 421

Moving on to evaluation….

How do we determine how well our classification models work?

When can we say that our performance is good?

When can we say that our model is better than others?

Gold Labels

- Before determining anything, we need some sort of basis upon which to make our comparisons
  - *Is "Sarcastic" the correct label for "Natalie told Usman she was soooo totally happy for him." ?*
- We can acquire **gold standard labels** from human annotators

Does it matter who our annotators are?

- Depends on the task
- For complex tasks, you may want to recruit experts in the desired subject area
  - Rating translation quality
  - Labeling pedagogical strategies in teacher-student interactions
- For simpler tasks, you can probably recruit non-experts
  - Deciding whether text is sarcastic or non-sarcastic
  - Deciding whether a specified event takes place before or after a second event
- Common sources of annotators:
  - Amazon Mechanical Turk: https://www.mturk.com
  - Appen: https://appen.com
  - Friends and family

Contingency Tables

- Once we have our gold standard labels (either from an existing dataset, or after collecting our own), we can begin comparing **predicted** and **actual** labels
- To do this, we can create a **contingency table**
  - Often also referred to as a **confusion matrix**

Contingency Tables

- In a contingency table, each cell labels a set of possible outcomes
- These outcomes are generally referred to as:
  - **True positives**
    - Predicted true and actually true
  - **False positives**
    - Predicted true and actually false
  - **True negatives**
    - Predicted false and actually false
  - **False negatives**
    - Predicted false and actually true

Predicted \ Actual		
	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

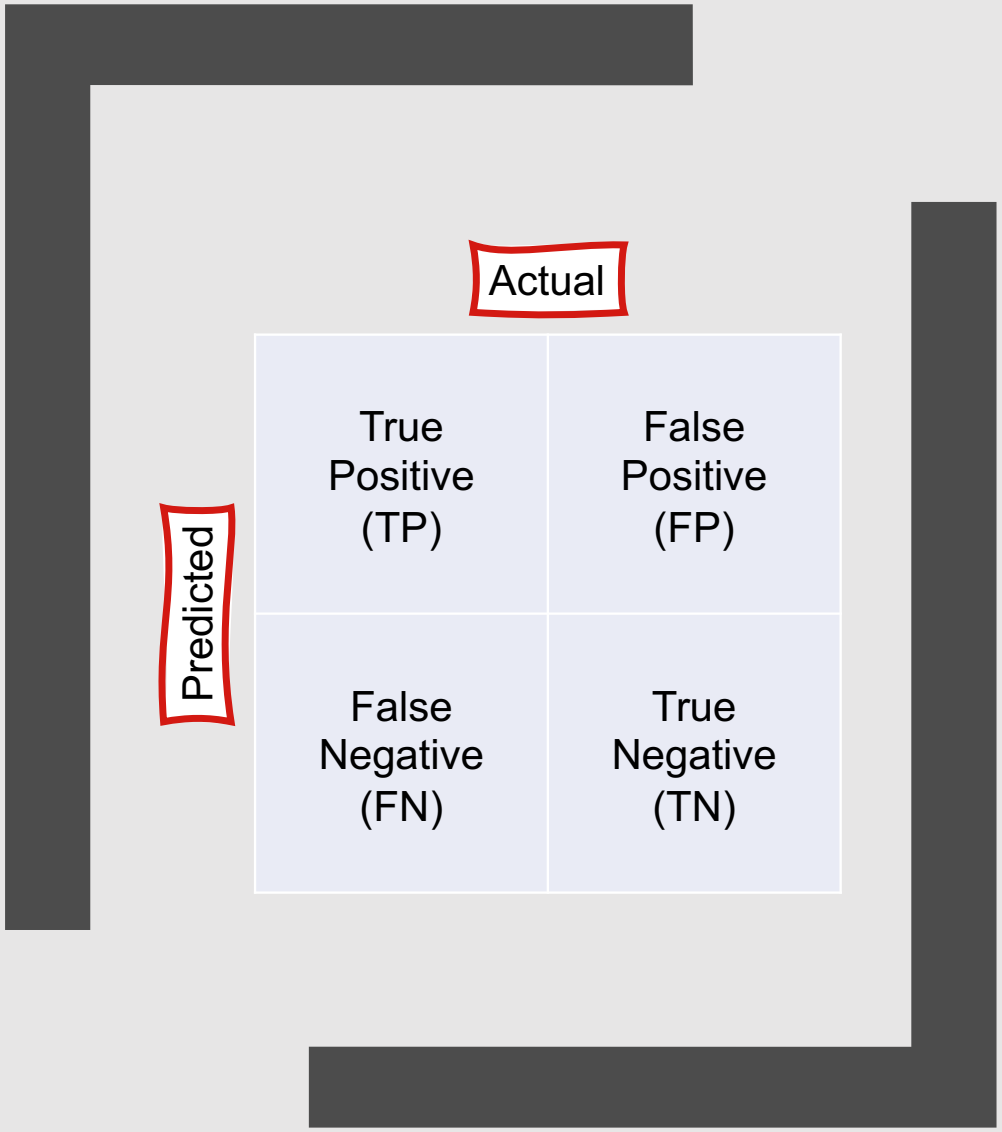
We can compute a variety of metrics using contingency tables.

Precision

Recall

F-Measure

Accuracy



Accuracy

- **Accuracy:** The percentage of all observations that the system labels correctly
- Accuracy = $\frac{tp+tn}{tp+fp+tn+fn}$

Why not just use accuracy and be done with it?

- This metric can be problematic when dealing with unbalanced datasets!
  - Imagine that we have 999,900 non-sarcastic sentences, and 100 sarcastic sentences
  - Our classifier might decide to just predict “non-sarcastic” every time to maximize its expected accuracy
    - 999900/1000000 = 99.99% accuracy
  - However, such a classifier would be useless …it would never tell us when a sentence *is* sarcastic

Thus, accuracy is a poor metric when the goal is to discover members of a less-frequent class.

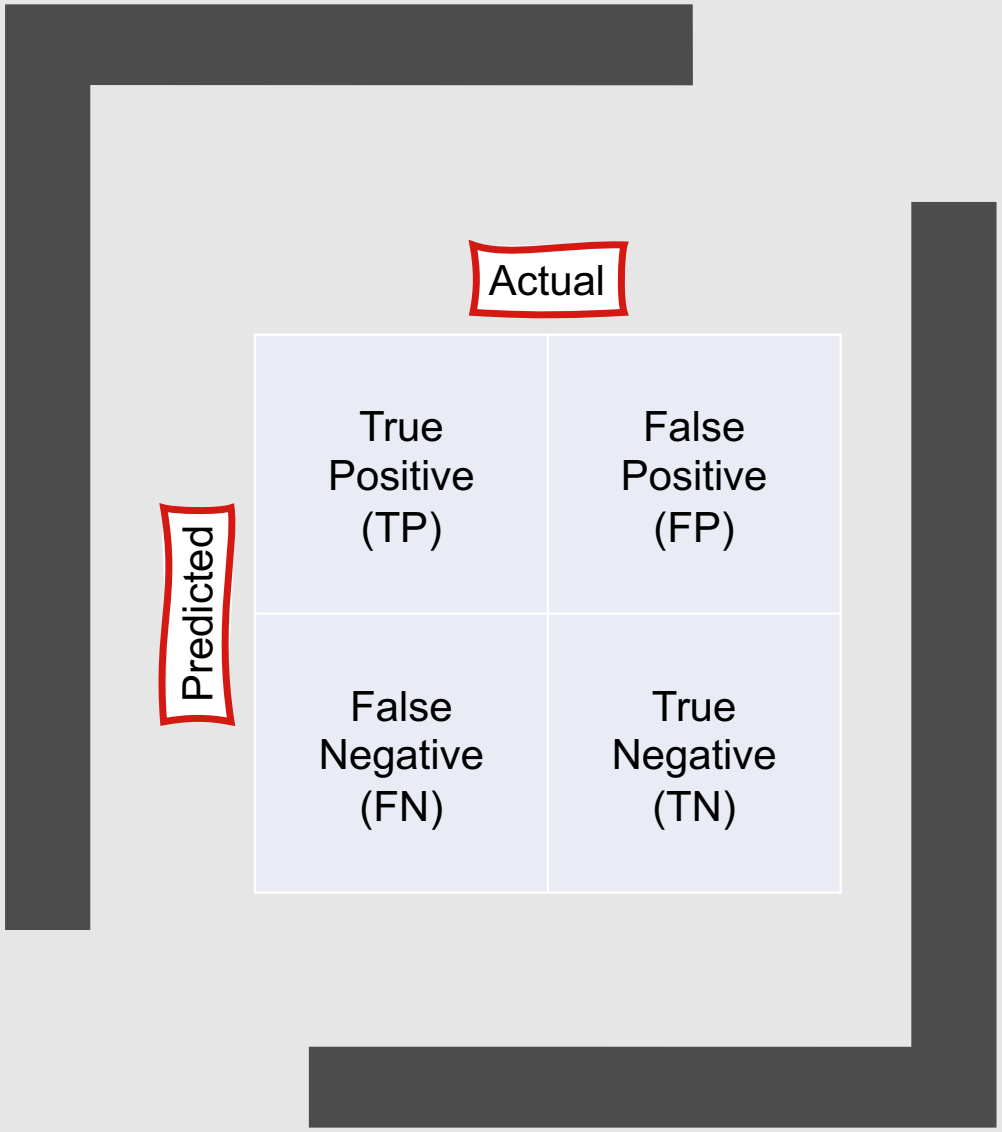
- Doing so is a very common situation
  - Detecting medical issues
  - Detecting papers dealing with a certain topic
  - Detecting spam

Precision

Recall

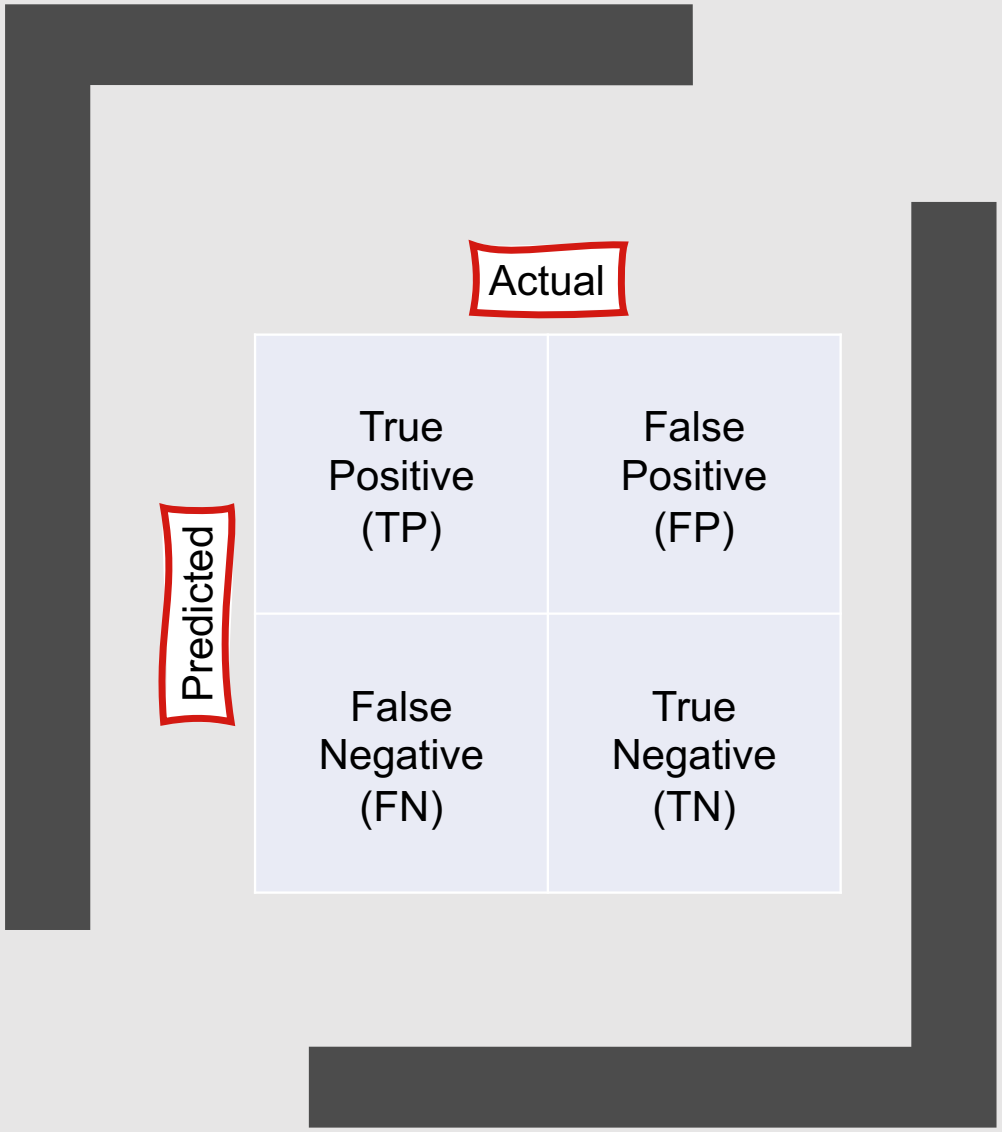
F-Measure

What are some alternatives that can focus on specific classes?



Precision

- **Precision:** Of the instances that the system predicted to be positive, what percentage actually are?
- Precision = $\frac{tp}{tp+fp}$



Recall

- **Recall:** Of the instances that actually are positive, what percentage did the system predict to be?
- Recall = $\frac{tp}{tp+fn}$

Precision and recall both emphasize a specific class of interest.

- Positive class can be whichever class you're interested in
  - **Sarcastic** or Non-Sarcastic
  - Positive or **Negative**
- Thus, in our problematic example case, precision and recall for the positive (sarcastic) case would both be 0
  - Precision = 0/(0+0) = 0
  - Recall = 0/(0+100) = 0

	Actual	
Predicted	TP: 0	FP: 0
	FN: 100	TN: 999,900

Which is more useful: Precision or recall?

- Depends on the task!
- If it's more important to maximize the chances that all predicted true values really are true, at the expense of predicting some of the true values as false, focus on precision
- If it's more important to maximize the chances that all true values are predicted to be true, at the expense of predicting some false values to be true as well, focus on recall

What if both are important?

- **F-measure** combines aspects of both **precision** and **recall** by computing their weighted harmonic mean
  - $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$
- The β parameter weights the importance of precision and recall, depending on the needs of the application
  - β > 1 means that recall is more important
  - β < 1 means that precision is more important
  - β = 1 means that the two are equally important

F-Measure

- Most commonly, researchers set β = 1 to weight precision and recall equally
- In this case, the metric is generally referred to as F_1
  - $F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R}$
- Although F-measure combines both precision and recall, it tends to be conservative; thus, the lower of the two numbers will factor more heavily into the final score

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	
Oh yay a five-hour Zoom meeting!!!	Sarcastic	
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	
When is the grocery store the emptiest?	Not Sarcastic	
I just love coronavirus.	Sarcastic	

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: ?	FP: ?
	FN: ?	TN: ?

Positive Class: Sarcastic

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 1	FP: ?
	FN: ?	TN: ?

Positive Class: Sarcastic

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 1	FP: 1
	FN: ?	TN: ?

Positive Class: Sarcastic

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 1	FP: 1
	FN: 3	TN: ?

Positive Class: Sarcastic

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 1	FP: 1
	FN: 3	TN: 2

Positive Class: Sarcastic

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 1	FP: 1
	FN: 3	TN: 2

Positive Class: Sarcastic

$$\text{Precision} = \frac{tp}{tp+fp} = \frac{1}{1+1} = 0.5$$

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 1	FP: 1
	FN: 3	TN: 2

Positive Class: Sarcastic

$$\text{Precision} = \frac{tp}{tp+fp} = \frac{1}{1+1} = 0.5$$

$$\text{Recall} = \frac{tp}{tp+fn} = \frac{1}{1+3} = 0.25$$

Example: Precision, Recall, and F_1

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 1	FP: 1
	FN: 3	TN: 2

Positive Class: Sarcastic

Precision = 0.5

Recall = 0.25

$$F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R} = \frac{2*0.5*0.25}{0.5+0.25} = 0.333$$

Example: Same, but what if the positive class is Not Sarcastic?

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

Predicted \ Actual		
	TP: ?	FP: ?
	FN: ?	TN: ?

Positive Class: Not Sarcastic

Precision = ?

Recall = ?

$$F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R} = ?$$

Example: Same, but what if the positive class is Not Sarcastic?

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

Predicted \ Actual		
	TP: 2	FP: 3
	FN: 1	TN: 1

Positive Class: Not Sarcastic

Precision = ?

Recall = ?

$$F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R} = ?$$

Example: Same, but what if the positive class is Not Sarcastic?

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

	Actual	
Predicted	TP: 2	FP: 3
	FN: 1	TN: 1

Positive Class: Not Sarcastic

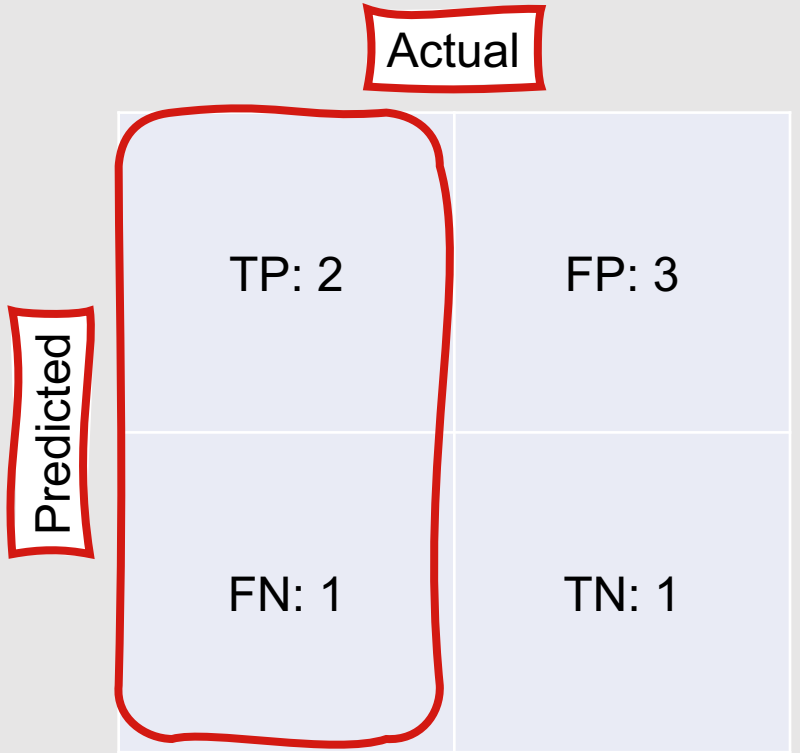
Precision = 0.4

Recall = ?

$$F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R} = ?$$

Example: Same, but what if the positive class is Not Sarcastic?

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic



Positive Class: Not Sarcastic

Precision = 0.4

Recall = 0.667

$$F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R} = ?$$

Example: Same, but what if the positive class is Not Sarcastic?

Instance	Actual Label	Predicted Label
I was absolutely thrilled that the pandemic was still ongoing.	Sarcastic	Not Sarcastic
I was absolutely thrilled that vaccine progress was going well!	Not Sarcastic	Not Sarcastic
Oh no I am soooo sad that my commute is only ten seconds.	Sarcastic	Sarcastic
Oh yay a five-hour Zoom meeting!!!	Sarcastic	Not Sarcastic
Oh yay my new order of hand sanitizer arrived!!!	Not Sarcastic	Sarcastic
When is the grocery store the emptiest?	Not Sarcastic	Not Sarcastic
I just love coronavirus.	Sarcastic	Not Sarcastic

Predicted \ Actual		
	TP: 2	FP: 3
	FN: 1	TN: 1

Positive Class: Not Sarcastic

Precision = 0.4

Recall = 0.667

$$F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R} = \frac{2*0.4*0.667}{0.4+0.667} = 0.50009$$