Supervised Learning for NLP II

HSS 510 / DS 518: NLP for HSS

Taegyoon Kim

Mar 25, 2025

Agenda

Things to be covered

- Various approaches to evaluation
  - Accuracy, precision, recall, F-1
  - ROC-AUC and PR-AUC
  - Metrics for continuous outcomes
- Determining a unit for manual labeling
- Allocating labeling resources
- Active learning for labeling text for imbalanced classification

Highlights from Last Week

Overview of text classification

- The goal is to
  - To classify documents into pre-defined categories
- We need to
  - Build a labeled data set (for training and testing)
  - Train a model to map texts to labels
  - Evaluate the model (using various performance metrics and tools like cross-validation)
- In a way that considers bias/variance trade-offs

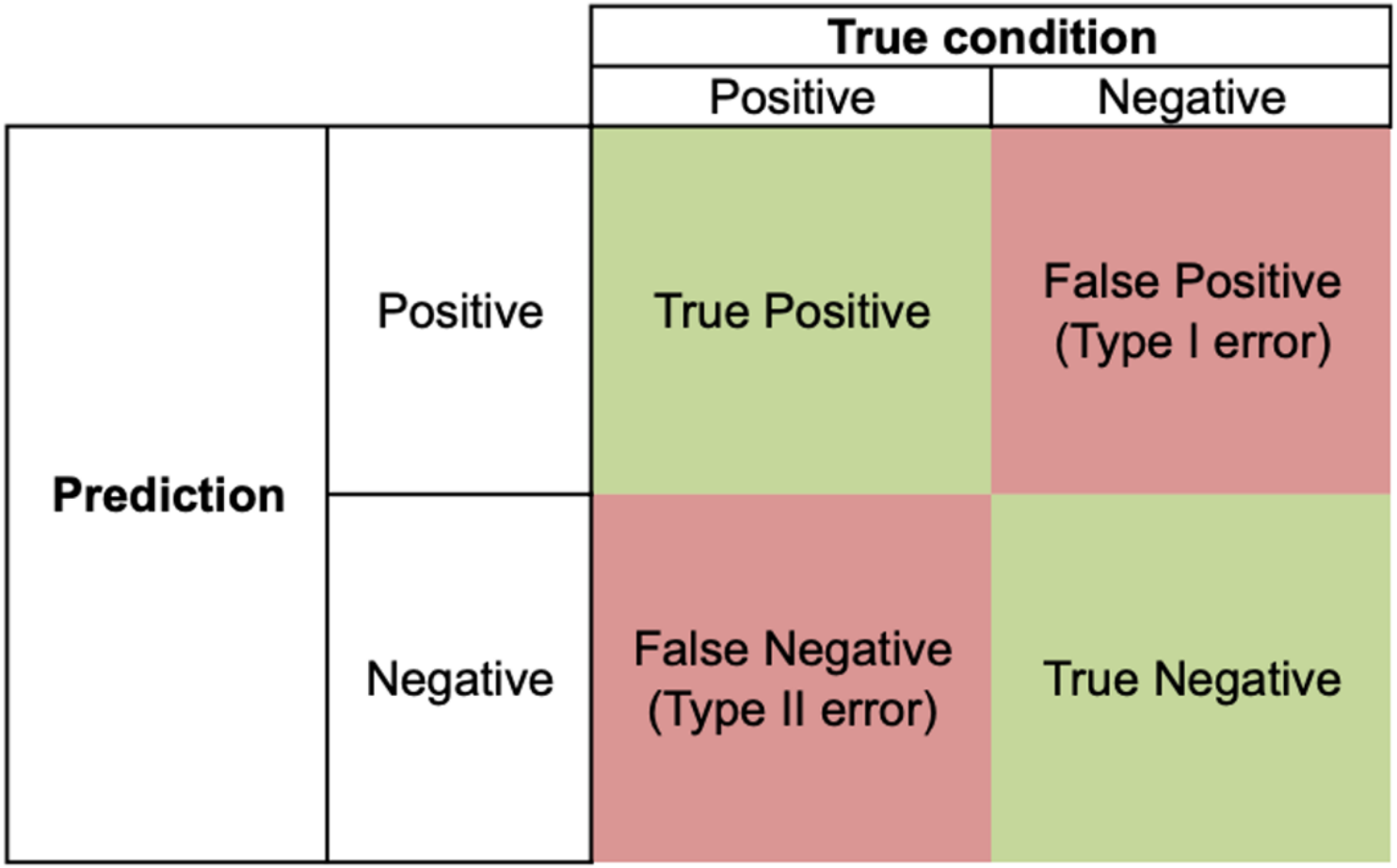
Various Approaches to Evaluation

Performance metrics

- Accuracy: the proportion of all predictions (both positive and negative) that the model got right
- Precision: the proportion of positive predictions that were actually correct
- Recall: the proportion of actual positives that were correctly predicted
- F-1: the harmonic (as opposed to arithmetic) mean of precision and recall

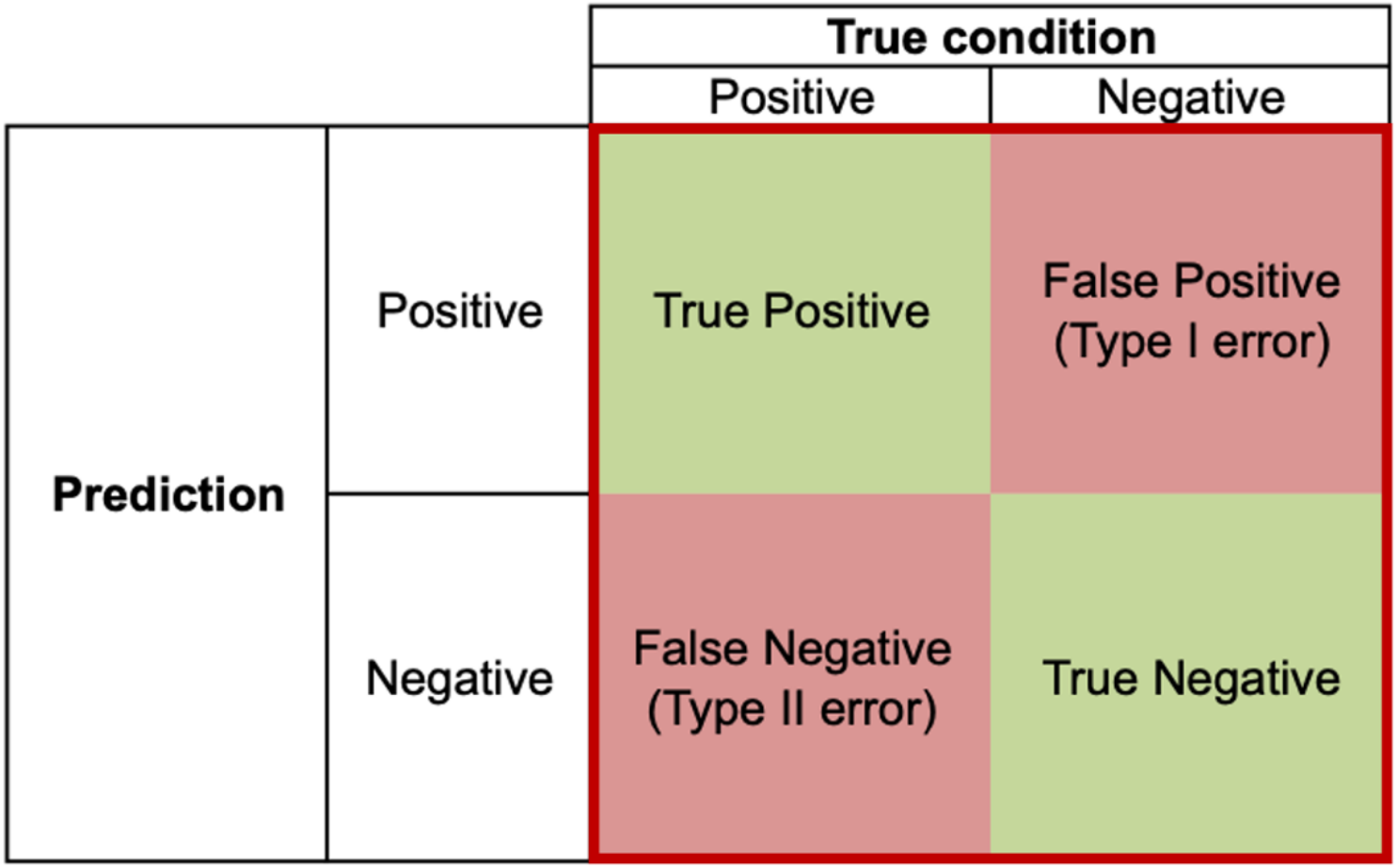
Various Approaches to Evaluation

Confusion matrix: predictions against true labels



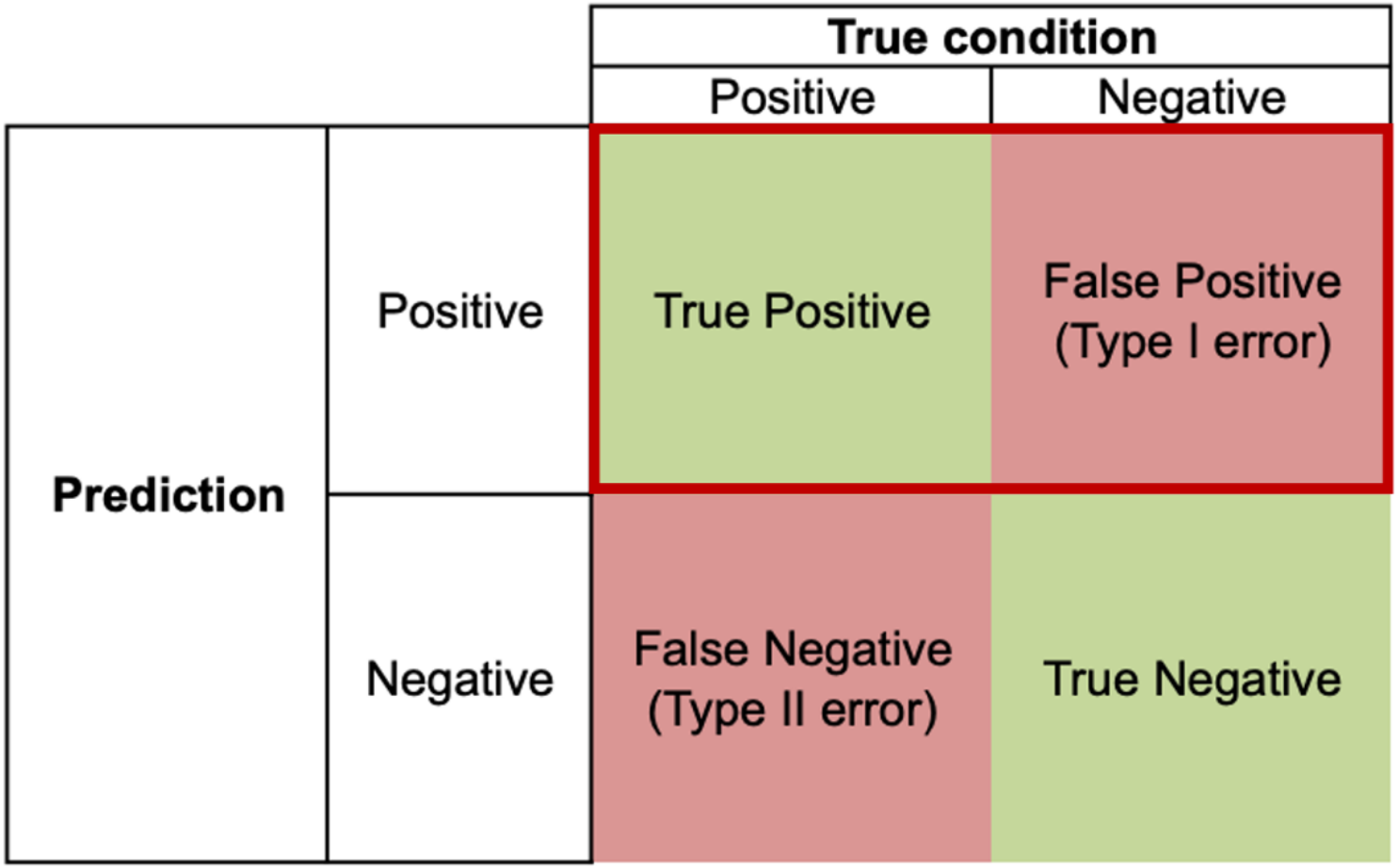
Various Approaches to Evaluation

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$



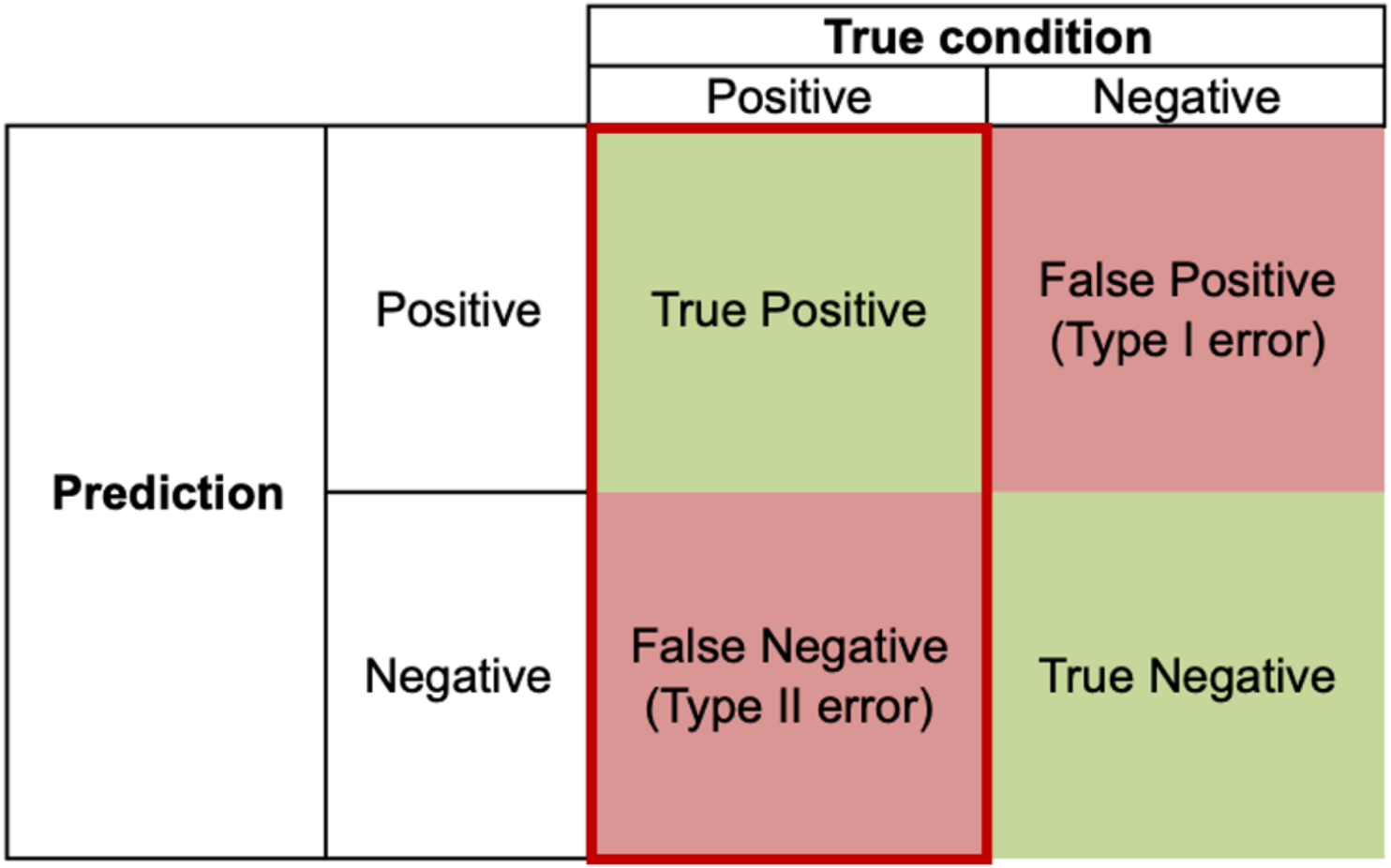
Various Approaches to Evaluation

Precision: $\frac{TP}{TP+FP}$



Various Approaches to Evaluation

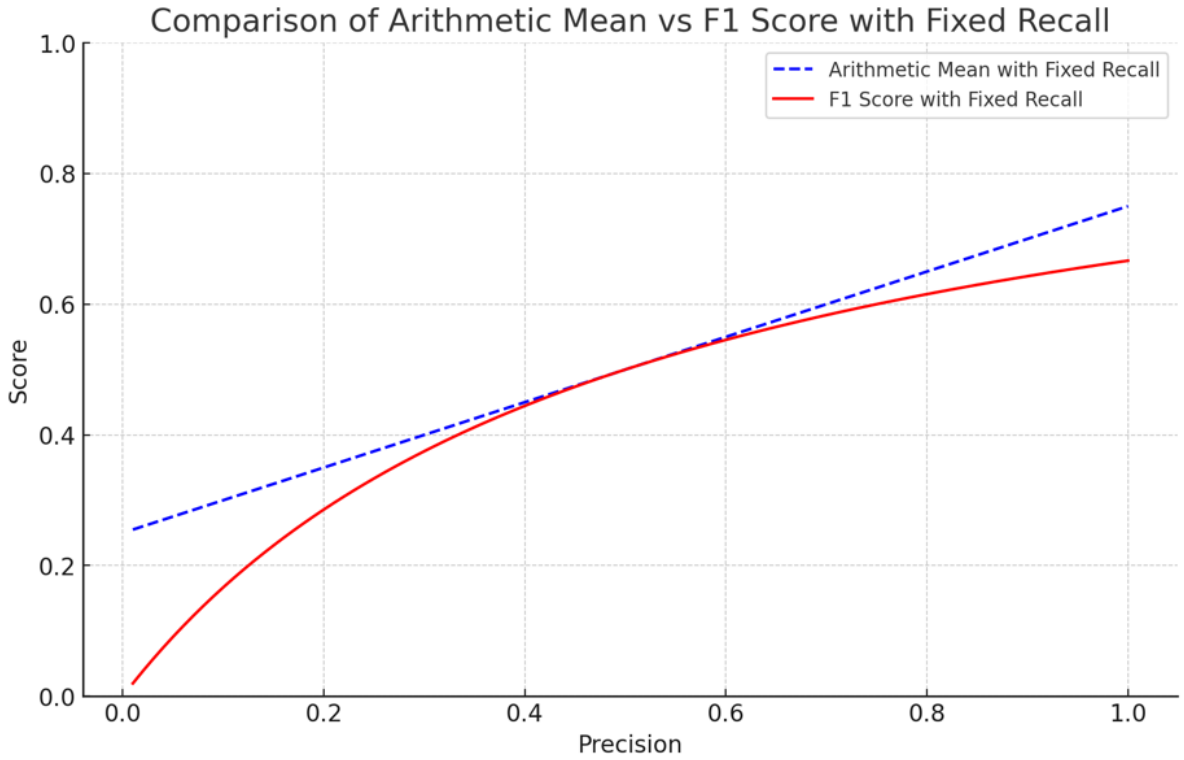
Recall: $\frac{TP}{TP+FN}$



Various Approaches to Evaluation

F-1: $(2 \times precision \times recall)$ / $(precision + recall)$

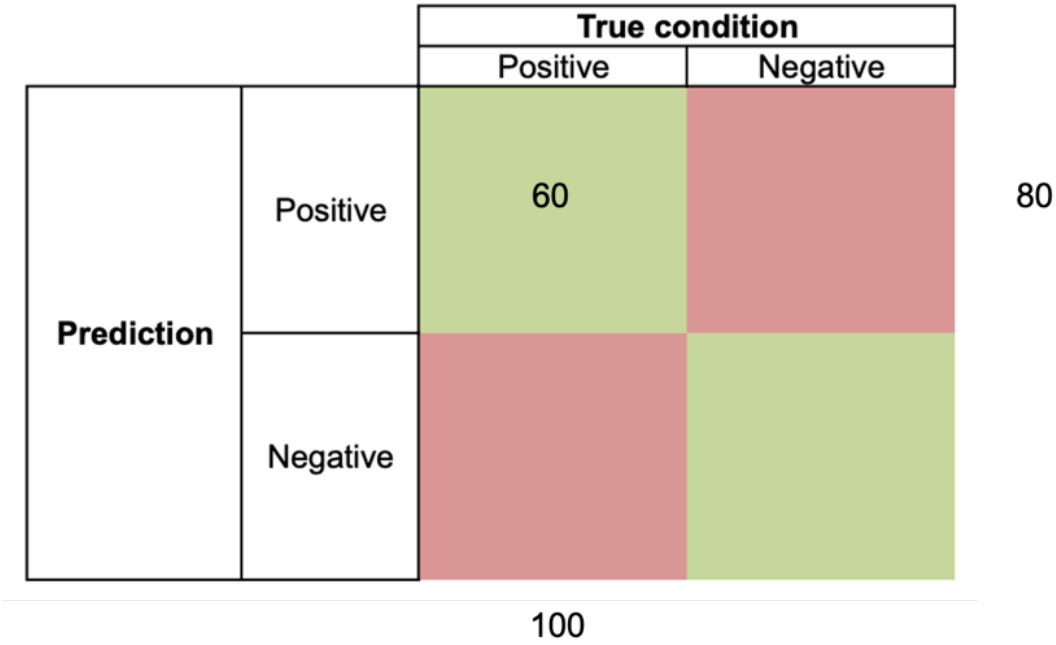
- Why not arithmetic mean $((precision + recall)/2)$?



Various Approaches to Evaluation

Precision/recall/F-1 & accuracy

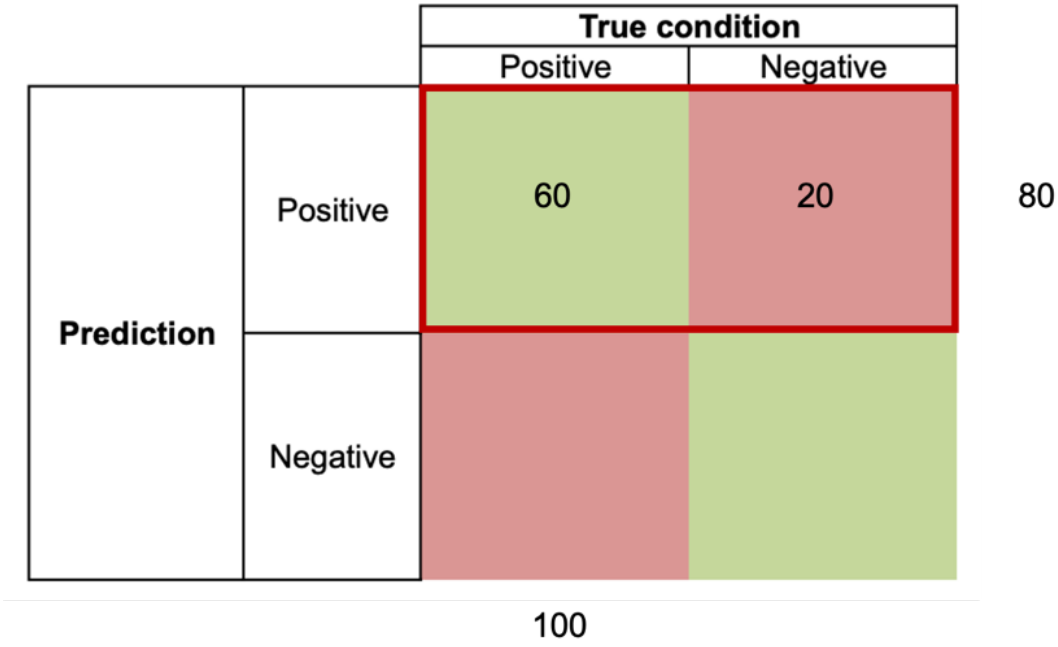
- 100 positives
- 80 predicted positives
- 60 true positives



Various Approaches to Evaluation

Precision/recall/F-1 & accuracy

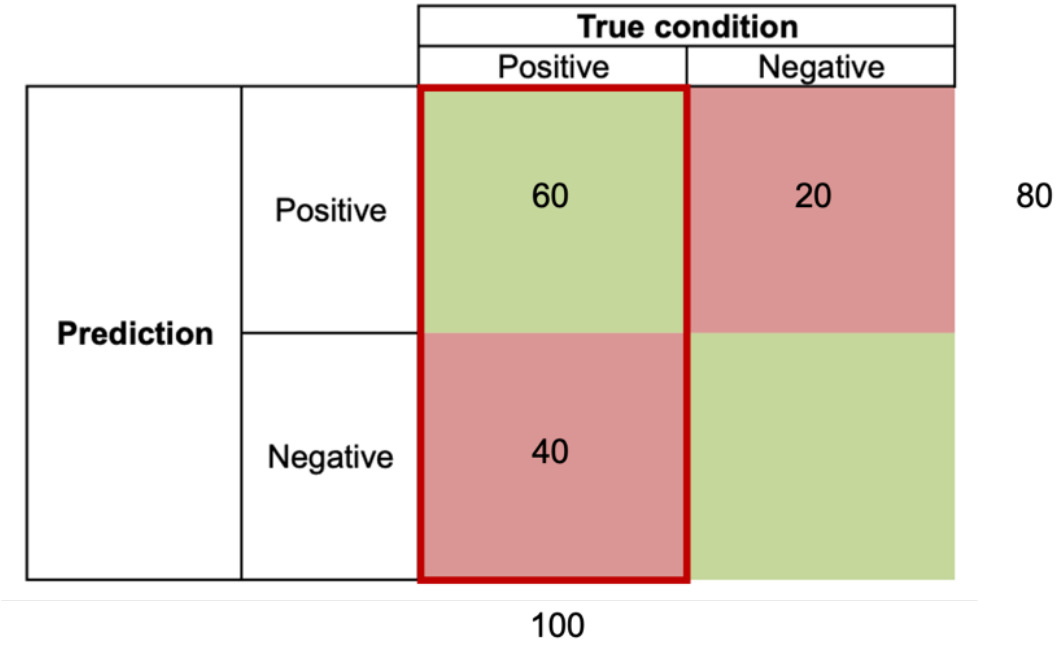
- Precision: $\frac{60}{60+20} = 0.75$



Various Approaches to Evaluation

Precision/recall/F-1 & accuracy

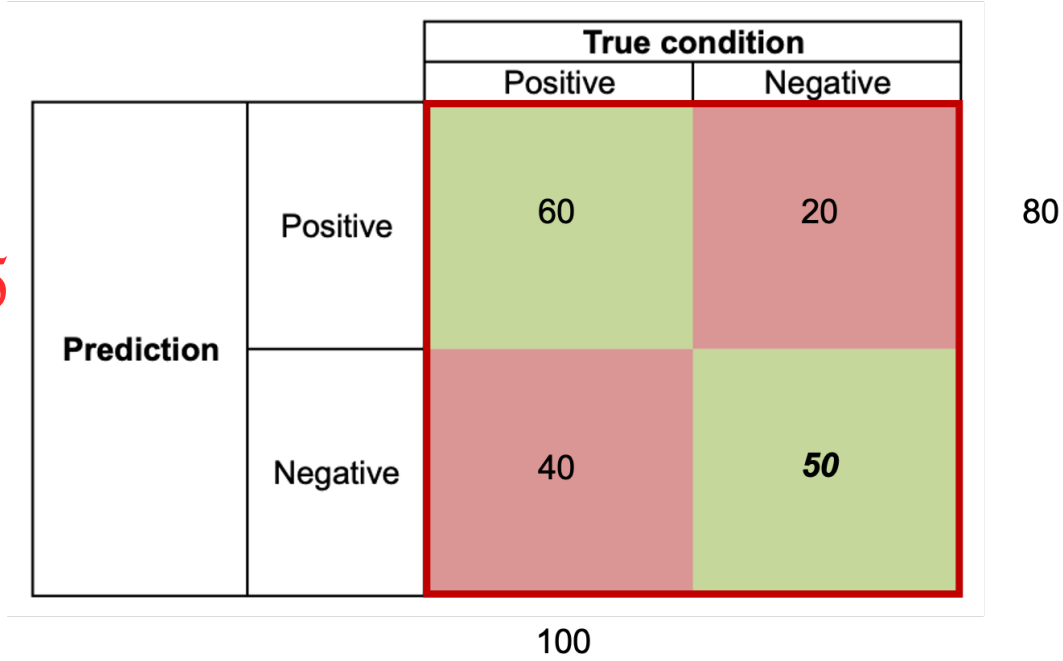
- Recall: $\frac{60}{60+40} = 0.6$



Various Approaches to Evaluation

Precision/recall/F-1 & accuracy

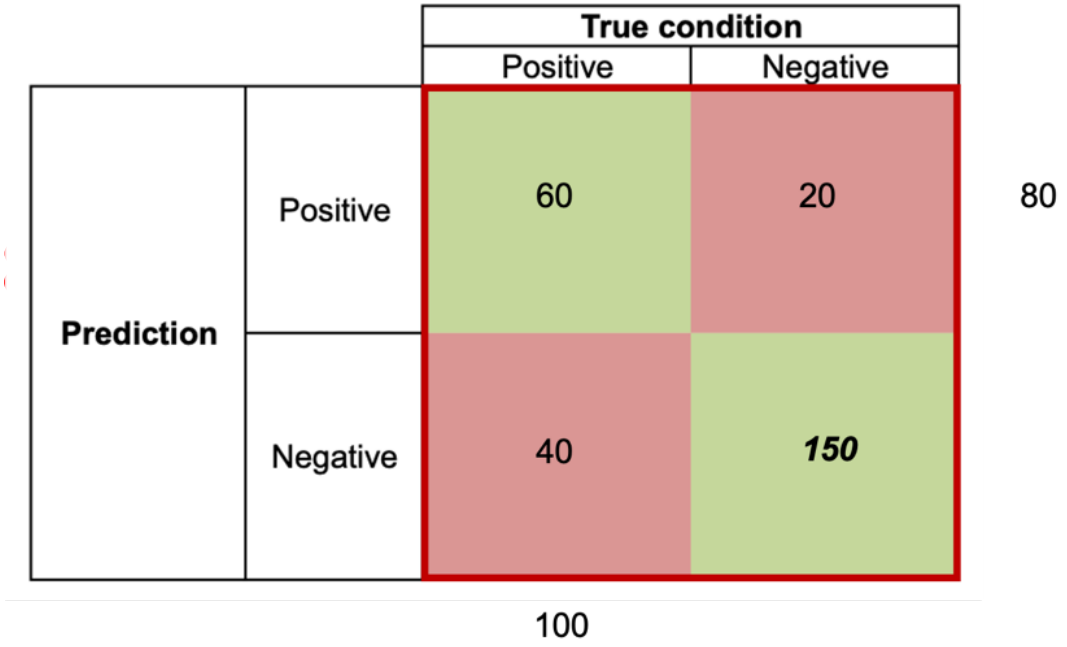
- Precision: $\frac{60}{60+20} = 0.75$
- Recall: $\frac{60}{60+40} = 0.6$
- Accuracy: $\frac{60+50}{60+20+40+50} = 0.65$



Various Approaches to Evaluation

Precision/recall/F-1 & accuracy

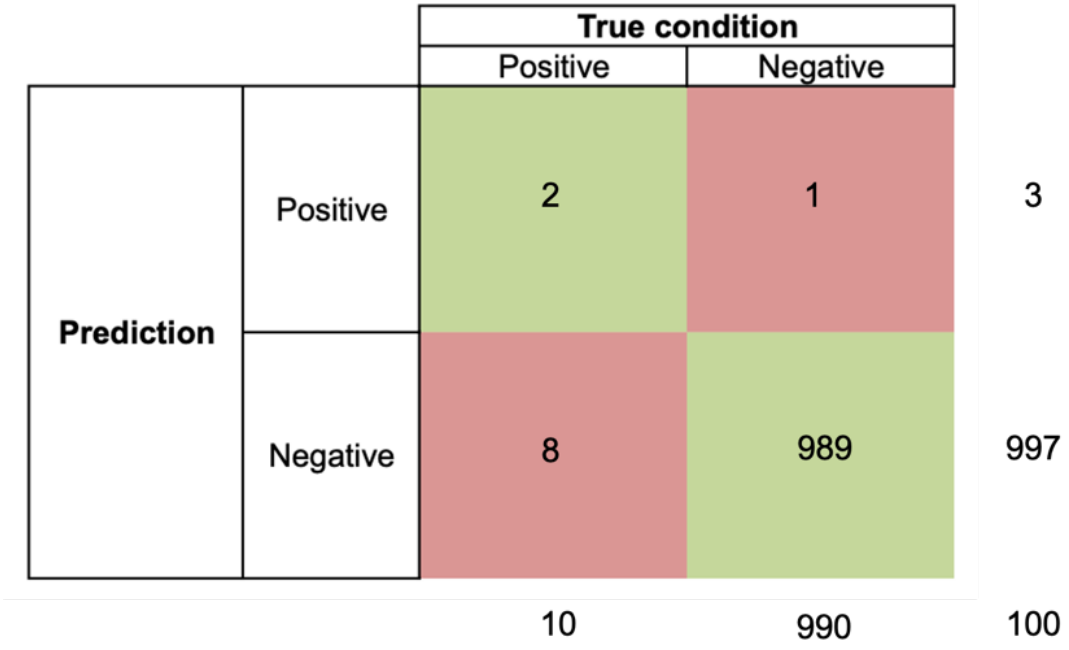
- Precision: $\frac{60}{60+20} = 0.75$
- Recall: $\frac{60}{60+40} = 0.6$
- Accuracy: $\frac{60+150}{60+20+40+150} = 0.7$



Various Approaches to Evaluation

An extremely imbalanced case

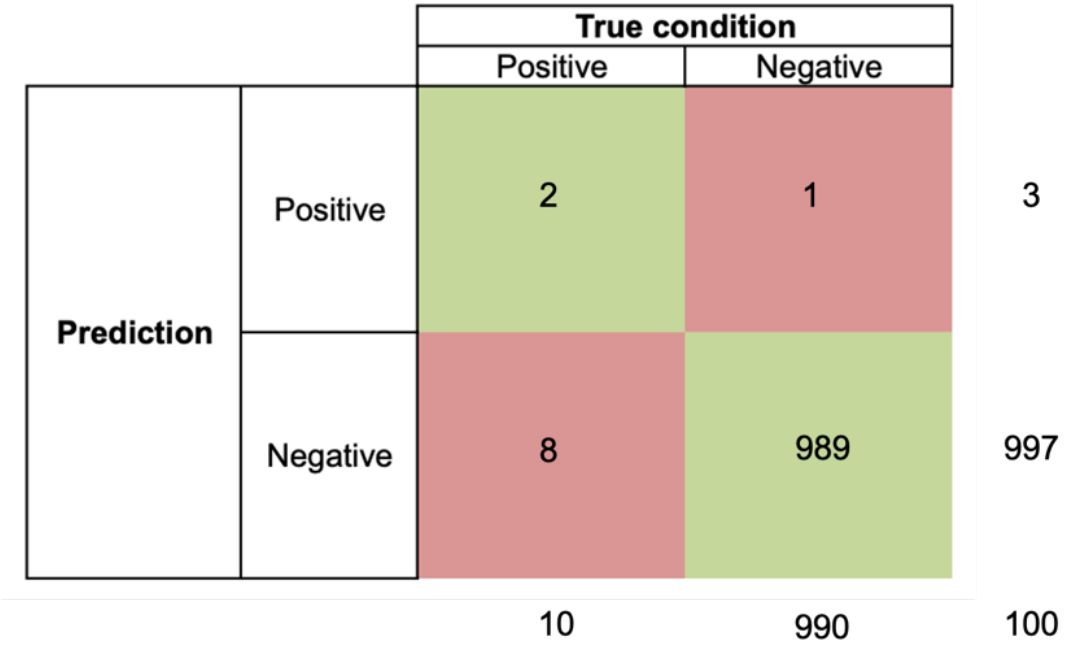
- Accuracy: ??
- Precision: ??
- Recall: ??
- F-1: ??



Various Approaches to Evaluation

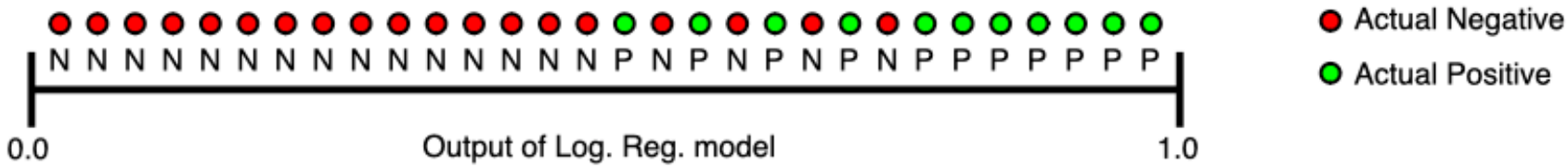
An extremely imbalanced case

- Accuracy: 0.991
- Precision: 0.66
- Recall: 0.2
- F-1: 0.31



Various Approaches to Evaluation

Accuracy, precision, recall, and F-1 are **static** metrics



- Calculated at a fixed decision threshold (typically 0.5 for binary classification)
- It might be useful to evaluate the model's performance across all possible thresholds
- This offers a more comprehensive evaluation

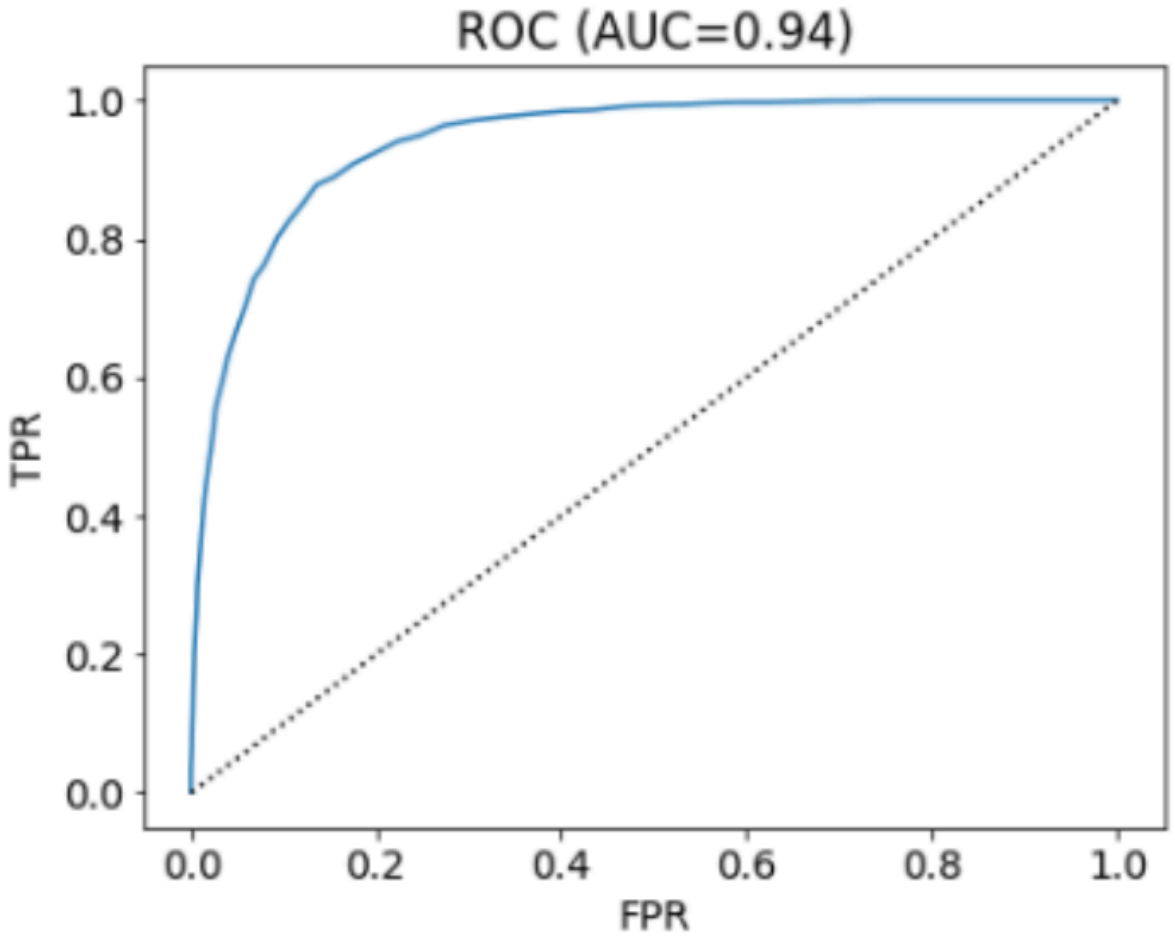
Various Approaches to Evaluation

ROC-AUC

- Receiver Operating Characteristic - Area Under the Curve
- TPR: $\frac{TP}{TP+FN}$ (= recall or sensitivity)
- FPR: $\frac{FP}{FP+TN}$
- Depicts the TPR (True Positive Rate) (y-axis) against FPR (False Positive Rate) (x-axis)
- Considers both positive (based on TPR) and negative classes (based on FPR)

Various Approaches to Evaluation

ROC-AUC



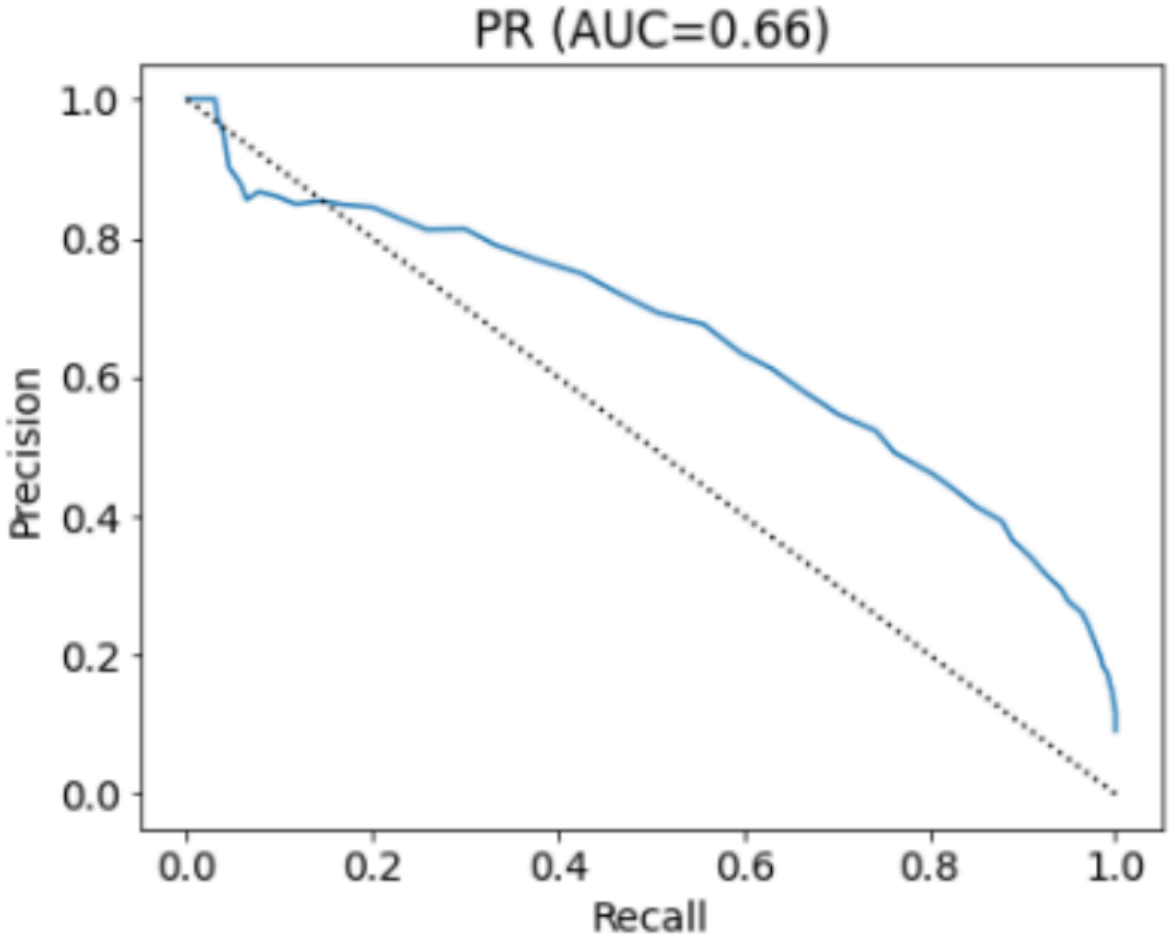
Various Approaches to Evaluation

PR-AUC

- Precision Recall Area Under Curve
- Depicts the precision against recall
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$
- Focused on positive class (look at the formulas for precision and recall, respectively)

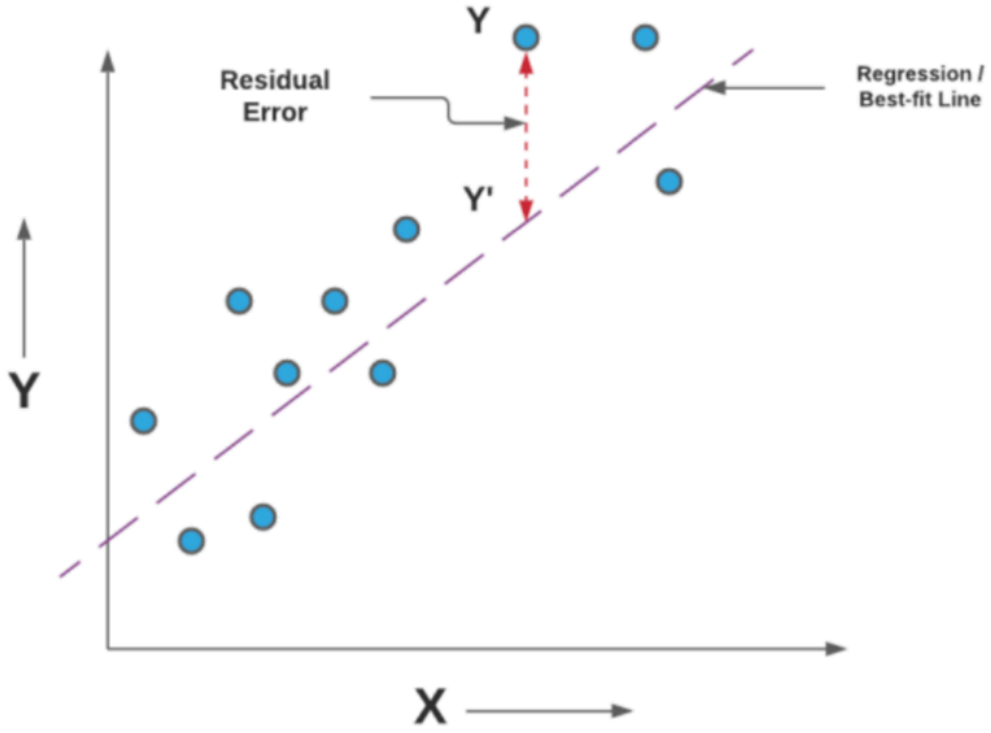
Various Approaches to Evaluation

PR-AUC



Various Approaches to Evaluation

How about continuous outcomes (e.g., OLS regression)?



$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

Determinig a Unit for Manual Labeling

What should the unit for labeling be?

- Text classification often focuses on a specific concept (e.g., sentiment, stance, or topic)
- Two considerations
  - In many cases, texts are lengthy, and individual sentences within a text might be irrelevant
  - Additionally, individual sentences within a text may vary in their alignment with the concept, adding further complexity
- See the next two example news articles about the U.S. economy in terms of sentiment

Determinig a Unit for Manual Labeling

Example Article I

The annual Riverfront Festival went ahead as planned this weekend despite heavy rain throughout the day.

Families turned out in large numbers, enjoying live music, food trucks, and craft stalls under tents and umbrellas.

Organizers praised the community’s resilience and confirmed plans to expand the event next year.

Some vendors, however, expressed concern that recent economic uncertainty in the U.S. had led to lower-than-expected sales.

Determinig a Unit for Manual Labeling

Example Article I

~~The annual Riverfront Festival went ahead as planned this weekend despite heavy rain throughout the day.~~

~~Families turned out in large numbers, enjoying live music, food trucks, and craft stalls under tents and umbrellas.~~

~~Organizers praised the community's resilience and confirmed plans to expand the event next year.~~

Some vendors, however, expressed concern that recent economic uncertainty in the U.S. had led to lower-than-expected sales.

Determinig a Unit for Manual Labeling

Example Article II

The U.S. economy continues to show resilience, with steady job growth and consumer spending remaining strong in recent months. Many businesses have reported increased sales, particularly in the retail and hospitality sectors.

However, economists warn that rising interest rates and persistent inflation are putting pressure on lower-income households, limiting their purchasing power.

Determinig a Unit for Manual Labeling

Example Article II

The U.S. economy continues to show resilience, with steady job growth and consumer spending remaining strong in recent months. Many businesses have reported increased sales, particularly in the retail and hospitality sectors. [positive]

However, economists warn that rising interest rates and persistent inflation are putting pressure on lower-income households, limiting their purchasing power. [negative]

Determinig a Unit for Manual Labeling

How to decide whether to label at the sentence level or at larger units such as paragraphs or entire texts?

1. Sentences irrelevant of the tone would only add noise to article-level classification
2. When the sentences in an article vary in terms of the concept, such variation will be lost when a single label is assigned

Determinig a Unit for Manual Labeling

Barbera et al. 2021

- Investigates whether sentence-level labeling is superior to labeling at larger chunks for article classification
- Train a classifier based on sentence-level labeling and a classifier based on segment-level labeling
- Compared their accuracy in predicting unseen segment-level ground truth data

Determinig a Unit for Manual Labeling

Barbera et al. 2021

- Task: sentiment analysis (positive vs. negative)
- Data: NYT articles about the U.S. national economy (1947–2014)
- Algorithm: logistic regression with an L2 penalty

Determinig a Unit for Manual Labeling

Barbera et al. 2021

- Accuracy scores: sentence-level classifier (0.7) vs. segment-level classifier (0.69)
- The choice of unit of labeling has little consequence
- There are few benefits in sentence-level labeling given additional resources spend on it

Determinig a Unit for Manual Labeling

Barbera et al. 2021

- Sentiments within a segment do not vary much
  - Among the set of positively-labeled segments, there is approximately 0.9 positive sentences and 0.27 negative sentences, on average
  - Among the set of negatively-labeled segments, there is approximately 1 negative sentence and 0.08 positive sentences, on average
- Segments include a substantial number of non-relevant sentences (2.64) (relative to relevant ones: 2.33)

Determinig a Unit for Manual Labeling

Takeaways

- We cannot overgeneralize these findings to other contexts
- However, the results provide evidence that sentence-level labeling offers little advantage when classifying texts in larger chunks
- Label by segment unless there is significant variation in tone across sentences and/or a high proportion of irrelevant sentences within segments

Allocating Labeling Resources

How should we allocate labeling resources?

- Texts are often (and ideally) labeled by multiple coders
- However, we have a limited number of coders and limited time
- We face a choice between (a) labeling more unique texts vs. (b) assigning more coders per document
- (a) results in a larger training set and thus higher performance
- (b) offers the opportunity to train coders and assess inter-coder reliability

Allocating Labeling Resources

Barbera et al. 2021

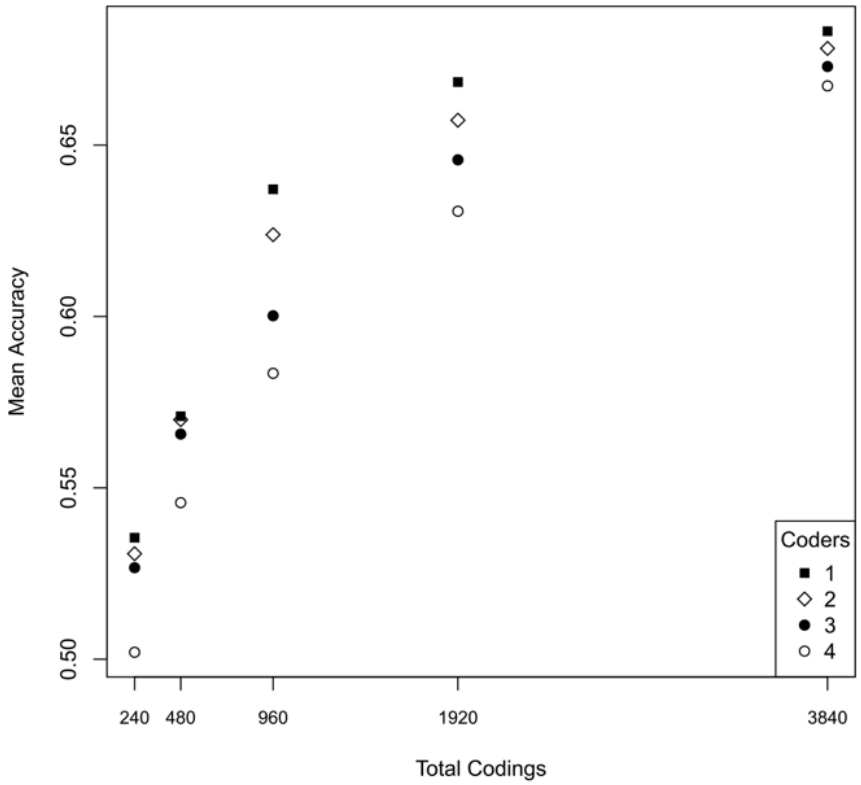


Figure 2. Accuracy with Constant Number of Total Codings.
Note: Results are based on simulations described in the text. Plotted points are jittered based on the difference from mean to clearly indicate ordering.

Allocating Labeling Resources

Difficult and subjective concepts (e.g., misogynistic speech)

- Poorly trained coders labeling unique texts can introduce bias and reduce performance
- Ensure thorough training using systematic labeling rules
- Ensure inter-coder reliability is assessed
- Once these steps are completed, focus on labeling as many unique texts as possible

Active Learning for Labeling Text

Cost of Manual labeling

- Expensive because labeled data is required to train a model
- Particularly expensive when classes are imbalanced
- A random sampling will yield a very small number of relevant documents
  - E.g., identifying news articles that refer to a terrorist attack
  - E.g., identifying social media posts containing hate speech
    - Siegel et al. (2019): less than 1% for most of the time

Active Learning for Labeling Text

Cost of Manual labeling

- Particularly relevant for data-hungry models (e.g., neural network-based models)
- Still relevant with transformer-based large language models
  - Model fine-tuning helps a lot, but does not eliminate the necessity for labeled data

Active Learning for Labeling Text

Active learning

- Choose data to be labeled in a way that reduces the labeling of documents
- Documents that are informative to the model are selected
- Increase in the efficiency of human labeling
- Meaningful with imbalanced classification
  - For balanced tasks, random sampling is likely to be effective

Active Learning for Labeling Text

Class imbalance

- One label occurs much more frequently than the other
- It takes an enormous amount of labeled data to get enough information about the minority class
- E.g., the relevant class appears 1% of the time
  - labeling of 5,000 documents → 50 relevant documents in expectation
  - The model trained on this data would likely be terrible

Active Learning for Labeling Text

For each iteration

- Label documents
- Train a a model
- Use the model to identify documents to label

New documents to label are selected based on the model's uncertainty ("margin sampling")

- Documents with predicted probabilities near 0.5 are ones that the model is least certain about

Active Learning for Labeling Text

Iterative process

1. Start with an initial set of labeled documents: $y_{labeled}$, $D_{labeled}$
2. Train a model F using $y_{labeled}$, $D_{labeled}$
3. Produce predicted probabilities for each unlabeled document in $D_{unlabeled}$
4. Select and label a new set of documents from $D_{unlabeled}$ based on model uncertainty ($|\hat{Pr}(y=1) - 0.5|$): y_z, D_z
5. Add y_z, D_z to $y_{labeled}$, $D_{labeled}$ (remove them from $D_{unlabeled}$)
6. Repeat Steps 1–5

Active Learning for Labeling Text

Experiments (Miller et al. 2020)

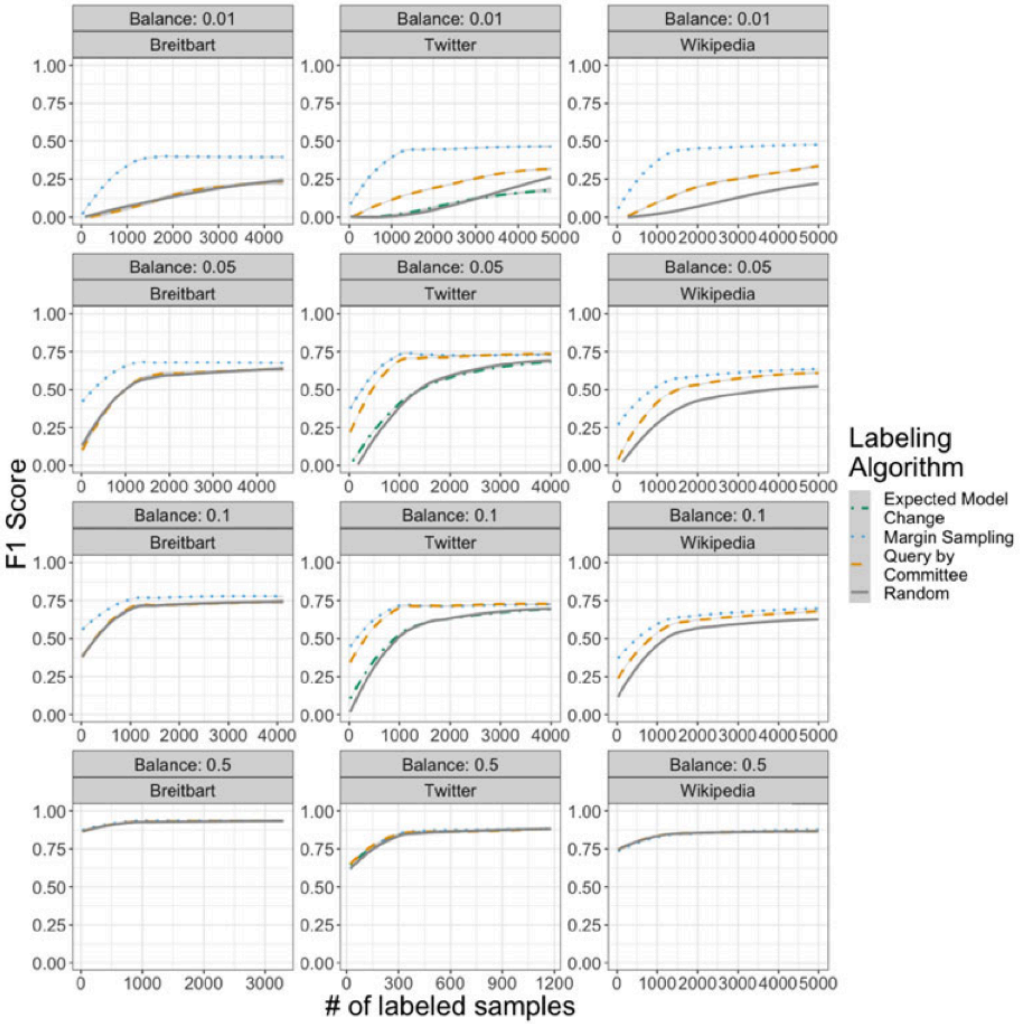


Figure 3. F1-score for experiments. The panel columns correspond to the datasets and the rows to the different levels of class imbalance. Dots represent single replications of the experiment and smoothed lines are fits (and standard errors) of a generalized additive model.

Active Learning for Labeling Text

Experiments (Miller et al. 2020)

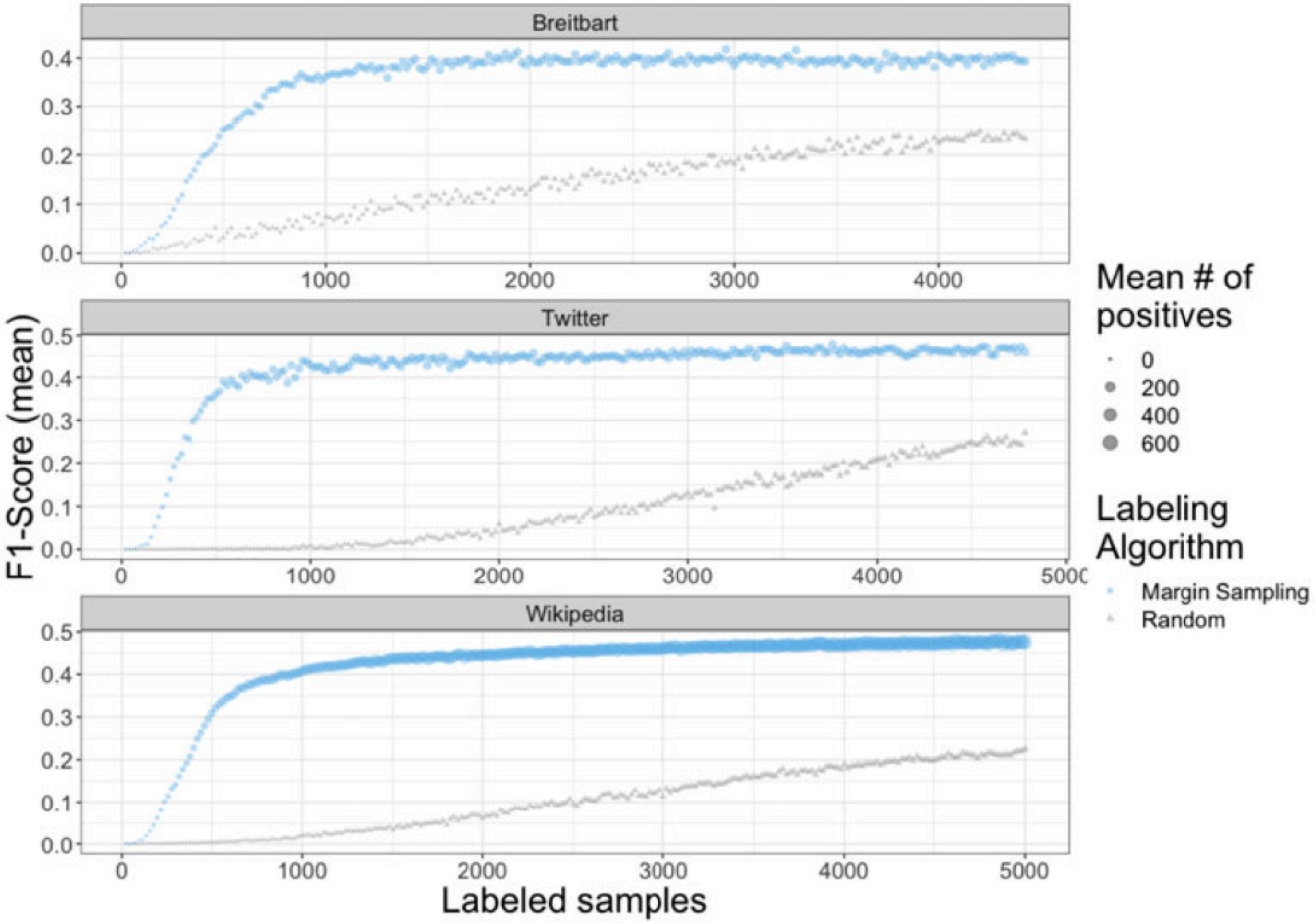


Figure 4. Performance by number of labeled examples for classifiers trained with active and passive learning (with class balance 0.01). Dots represent the average classifier performance across replications. Dot size is proportional to the average number of positively labeled observations in the training sample across replications.

Active Learning for Labeling Text

Experiments (Miller et al. 2020)

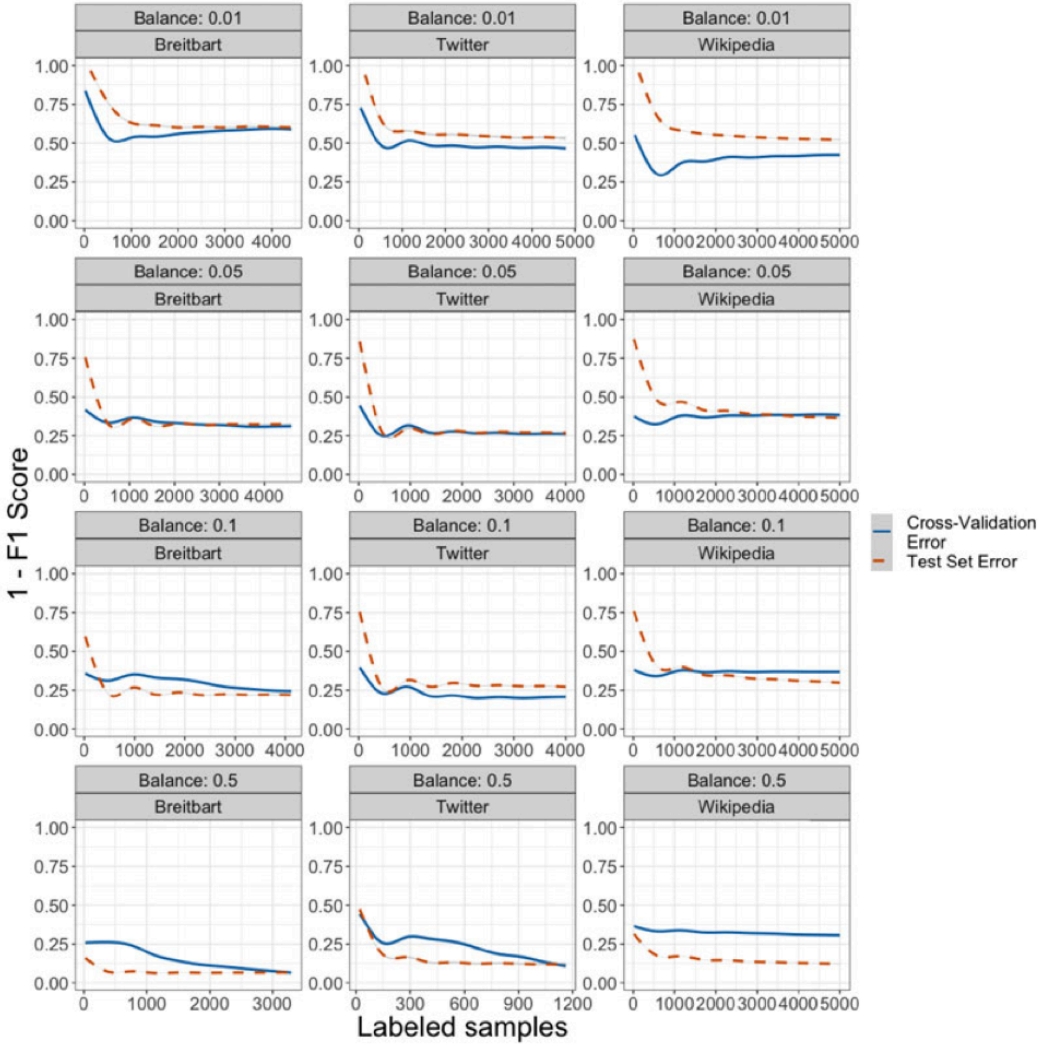


Figure 6. Comparison of generalization error (in terms of F1-score) in training and test set for active learning with margin querying strategy. Results are from the Twitter dataset.

Summary

Text classification with supervised learning

- Importance of carefully generated labeled data
  - Clear conceptualizations and labeling rules, assessing ICR, etc.
- Determining the unit for labeling
- Considerations regarding the resources dedicated to generating annotations
  - In general, having more unique labeled texts is preferable
  - For imbalanced classification, active learning can reduce the time and labor burden associated with manual labeling

Guided Coding

Identifying violent threats from Twitter by fine-tuning BERT (link)