precision and recall

André Santos, afs@inesctec.pt

Common NLP tasks

- paragraph/sentence splitting
- named entity recognition
- removing/restoring diacritics
- …

How to evaluate results?

Evaluation

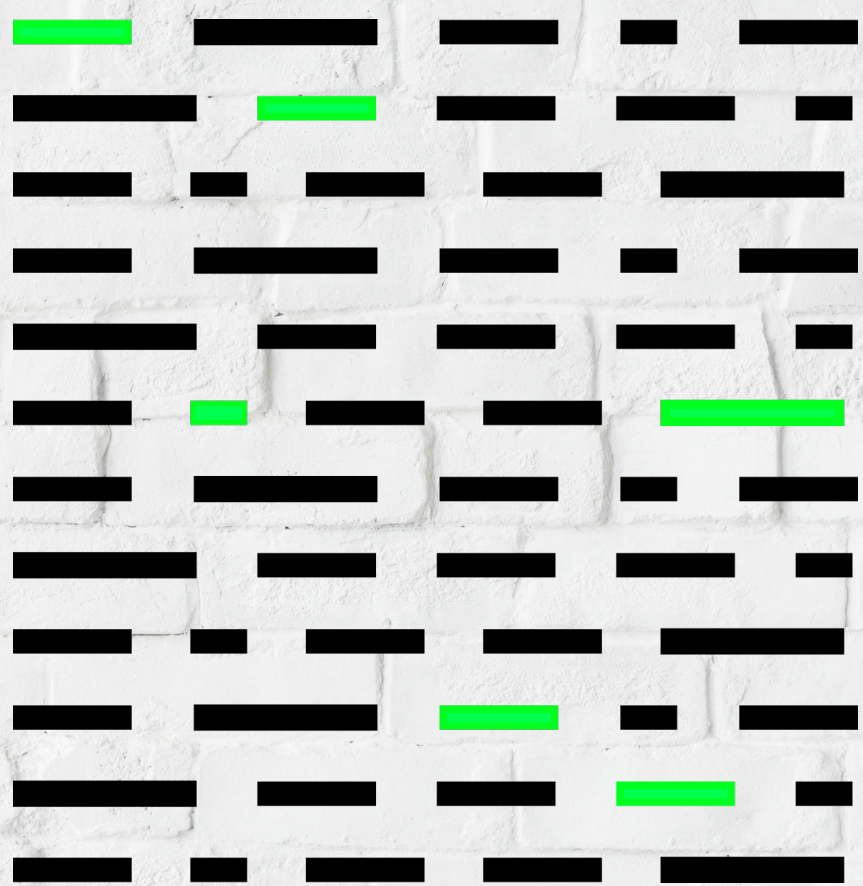
- **Comparison with a reference dataset (for example, a manually annotated collection of texts)**
- **When developing this kind of applications, dataset is often divided in (at least), two subsets:**
  - **train**
  - **evaluate**

Example

- Find first names in text
- For each word, determine whether it is an first name or not

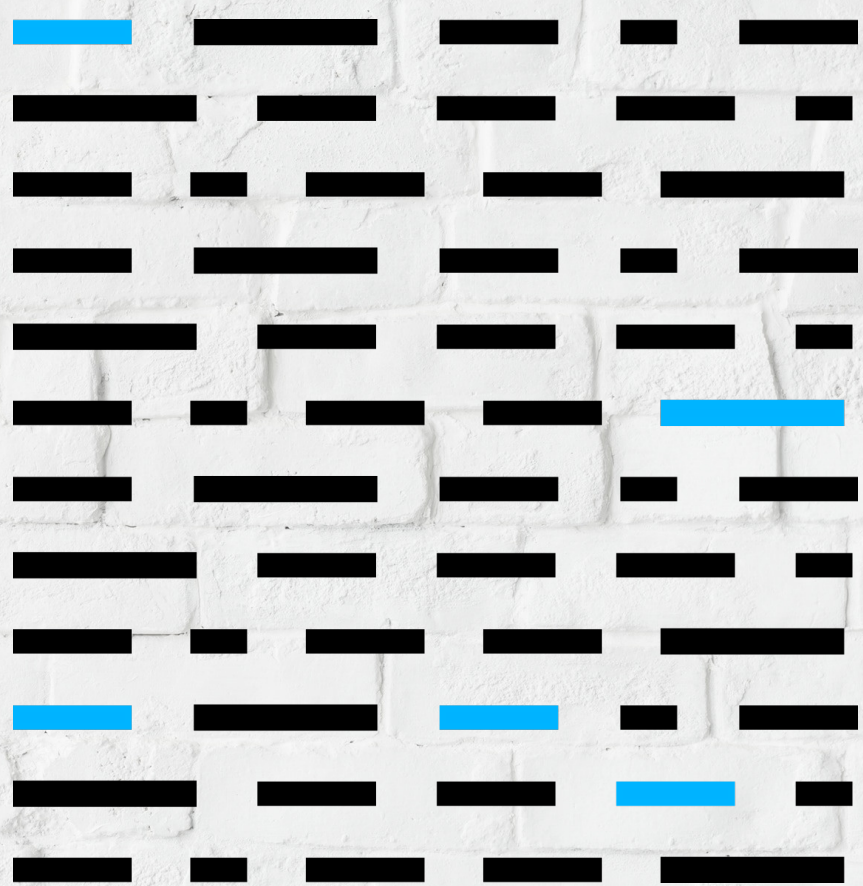
Example

- Let's consider a manually annotated text
- 60 words
- 6 first names

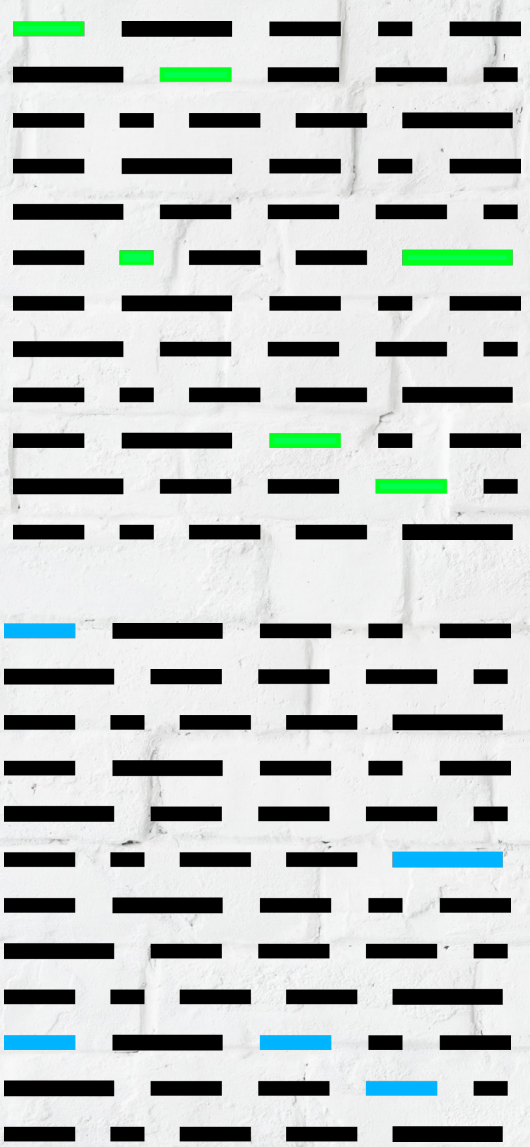


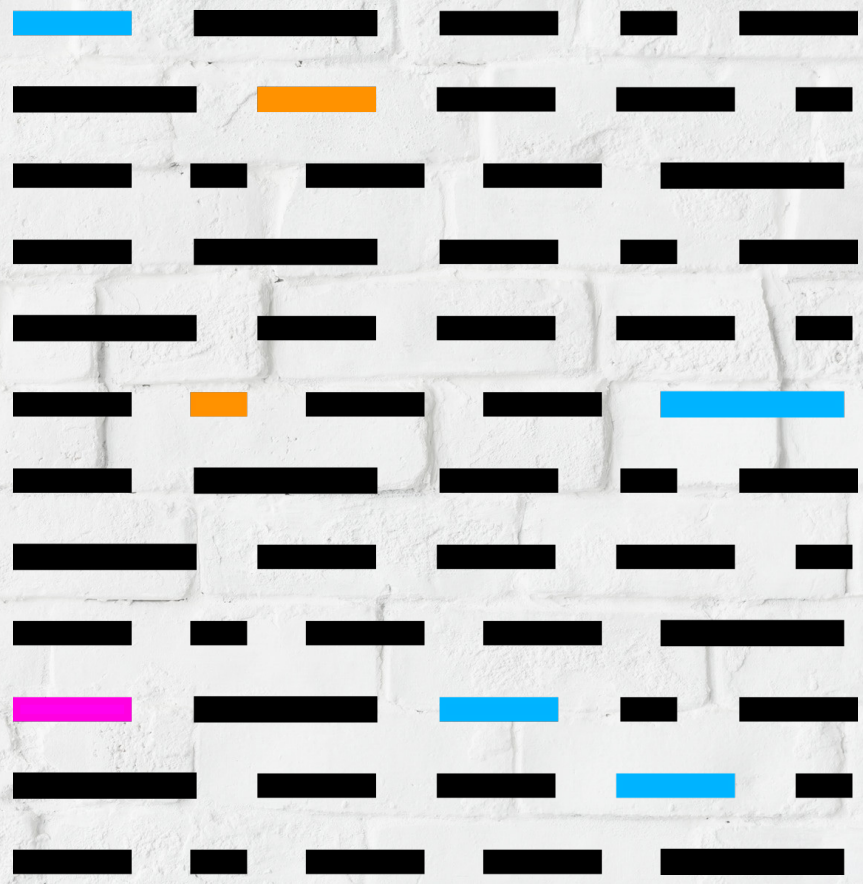
Example

- And our tool's output
- 6 first names, but not exactly the same 6 names



Example





Accuracy

- **percentage of correctly classified words**
- $$\frac{\text{number of correctly classified words}}{\text{total number of words}}$$
- **57 / 60 = 0.95 = 95%**

Confusion matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Confusion matrix

		Actual	
		First name	Not first name
Predicted	First name	4	1
	Not first name	2	54

Precision

- **represents the proportion of first names correcly classified**
- $$\frac{\texttt{true positives}}{\texttt{true positives + false positives}}$$
- $$\texttt{precision} = \frac{4}{4 + 1} = 0.8$$

Recall

- **expresses the ability to find all the relevant instances**
- $$\frac{\texttt{true positives}}{\texttt{true positives + false negatives}}$$
- $$\texttt{recall} = \frac{4}{4 + 2} = 0.667$$

F_1 score

- **harmonic mean between precision and recall**
- $$F_1 = 2 * \frac{precision * recall}{precision + recall}$$
- $$F_1 = 2 * \frac{0.8 * 0.667}{0.8 + 0.667} = 0.727$$

Trade-off between precision and recall

- **Classifiers often provide likelyhood values:**
  - **“I’m 8/10 sure Bruno is a first name”**
- **Trade-off between precision and recall can be adjusted by changing the cut-off value:**
  - **“Consider first names entities with a likelyhood value over 0.7”**

TPR and FPR

True Positive Rate

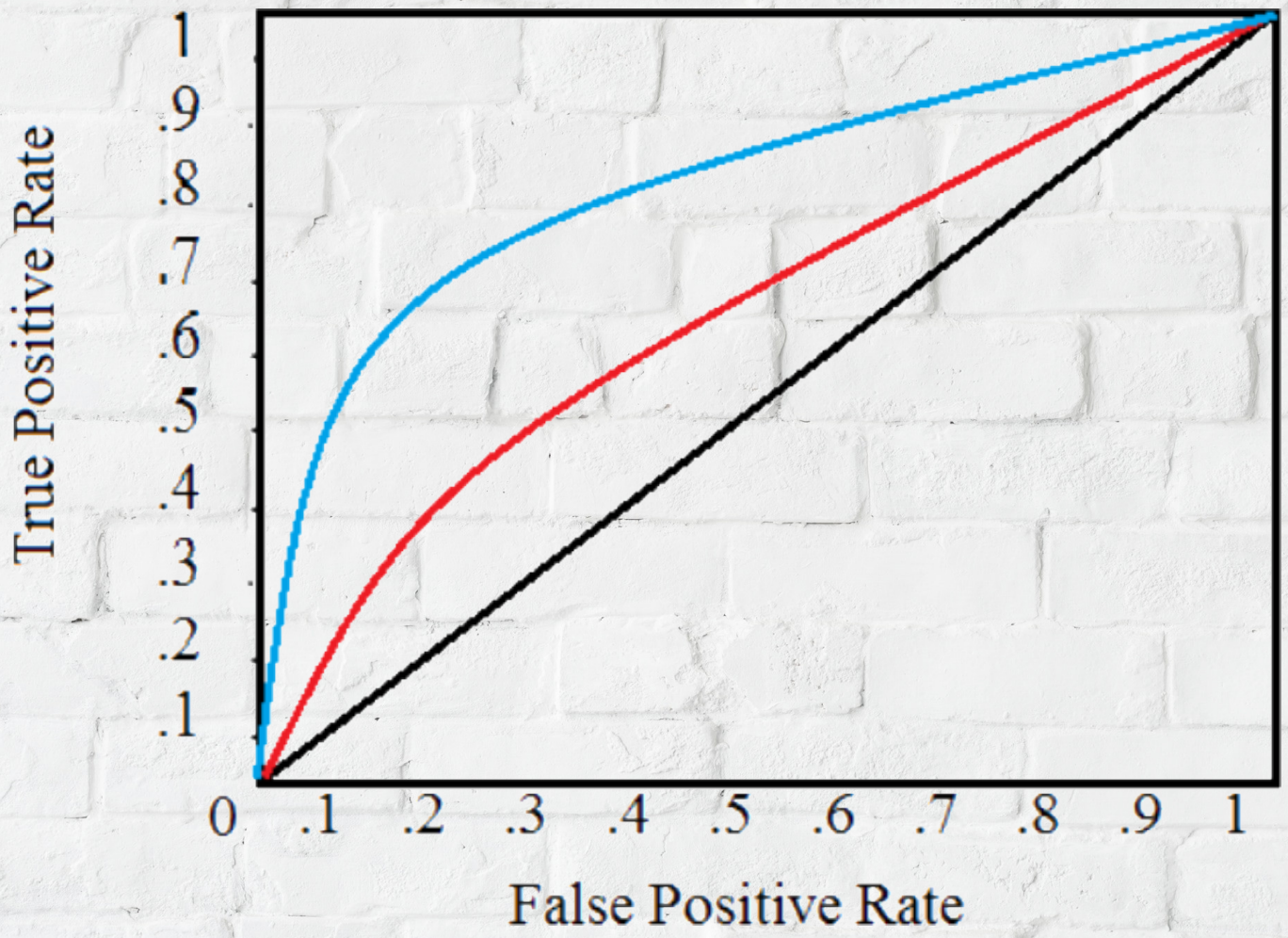
$$\frac{\text{true positives}}{\text{true positives + false negatives}}$$

False Positive Rate

$$\frac{\text{false positives}}{\text{false positives + true negatives}}$$

ROC curve

- **Receiver Operating Characteristic curve**



AUC

- **Area Under the Curve**
- **measures the area below a ROC curve**
- **allows comparisons which are independent of the cut-off points chosen**

Limitations

- **Unavailability of reference datasets**
  - **precision might still be measurable**
- **Sometimes, it is not clear what can be considered a false positive or false negative**