GEI1002

Computers and the humanities

Week 6
Working with text, Part II

NLP for more complex tasks

- Providing the summary of a document
- Finding the main themes or claims in a text
- Classifying sentences, paragraphs or whole texts

Classifying texts

Examples of possible tasks for literary texts

1. Literary genre classification
2. Theme identification
3. Period classification
4. Cultural context classification
5. Writing style analysis
6. Tone/mood classification
7. Poetic structure analysis

Classifying texts

Examples of possible tasks for news articles

1. Factual statement vs. opinion
2. News event vs. background information
3. Direct quotation vs. paraphrase
4. Attribution vs. non-attribution
5. Positive vs. negative sentiment
6. Main topic vs. sub-topic
7. Cause vs. effect
8. Temporal classification
9. Call-to-action vs. informative
10. Eyewitness account vs expert statement

Two challenges

Technical

Interpretive

Our goal: label assignment

Sentence 1 ← Label 1

Sentence 2 ← Label 1

Sentence 3 ← Label 2

Sentence 4 ← Label 2

Sentence 5 ← Label 1

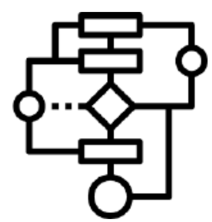
Using machine learning (ML)

Supervised Machine Learning

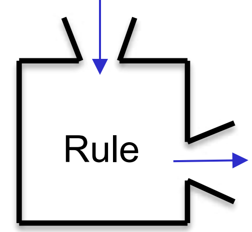
Data → Training → Model

Training set

- Sentence 1 Label 1
- Sentence 2 Label 1
- Sentence 3 Label 2
- Sentence 4 Label 2
- Sentence 5 Label 1

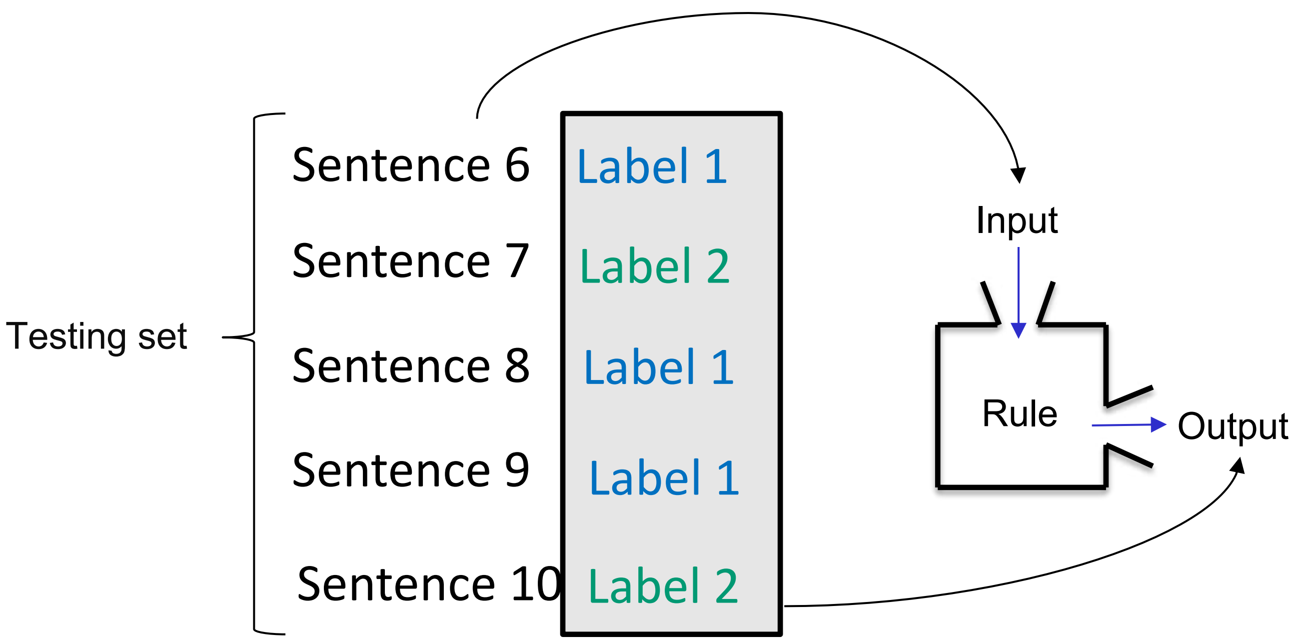


Input



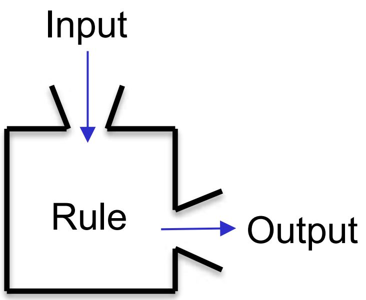
Output

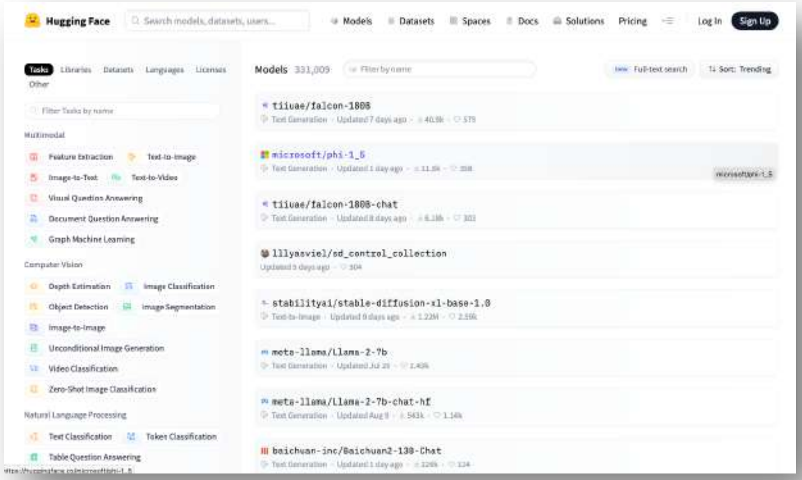
Model evaluation



Using machine learning (ML)

Once someone has trained a model, other people can use it

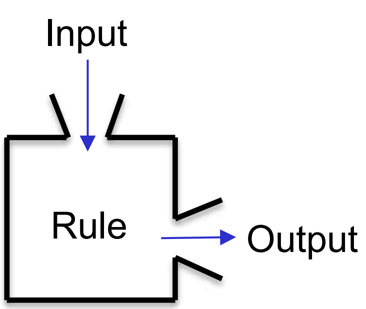




You can find a list of free, open-source models at https://huggingface.co/models

Using machine learning (ML)

Let’s consider an example



Model: distilbert-base-uncased-finetuned-sst-2-english

Today is a nice day. Positive

Today is a sad day. Negative

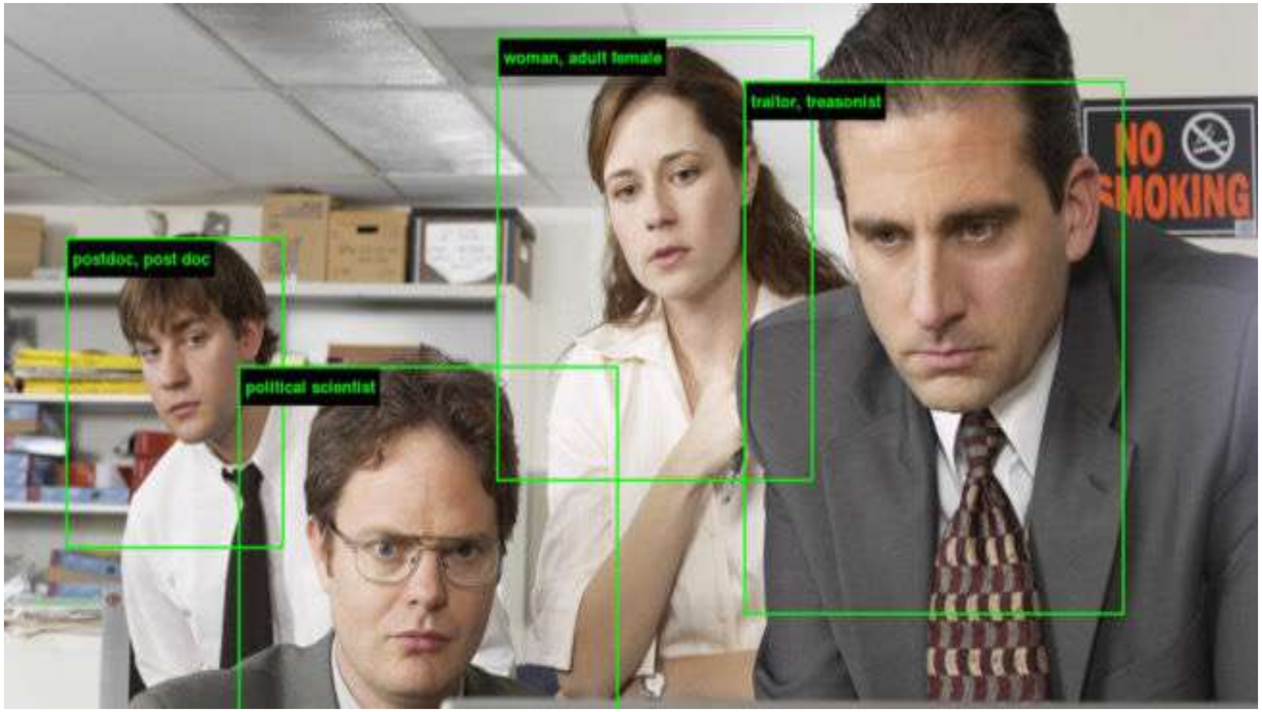
Today is a Tuesday. Positive ???

Using machine learning (ML)

A model is only as good as the data it was trained on. This includes the training labels.

Using machine learning (ML)

ImageNet Roulette by Trevor Paglen and Kate Crawford



https://paglen.studio/2020/04/29/imagenet-roulette/

No dataset is value-neutral

“Every dataset used to train machine learning systems, whether in the context of supervised or unsupervised machine learning, whether it's seen to be technically biased or not, **contains a worldview**. To create a training set is to take an almost infinitely complex and varied world and fix it into taxonomies composed of discrete classifications of individual data points, a process that requires inherently political, cultural, and social choices. By paying attention to these classifications, we can glimpse the various forms of power that are built into the architectures of AI world-building”

(Crawford, 2021, 133)

intricacies of classification - and the role tech plays

Where do labels come from?

- “Training” means building a set of rules (or “model”) from a set of labelled examples (“training set”).
- Where do these labels come from? Typically, from human labelers that assign them to thousands or millions of data points.
- This is often called ‘ground truth’. But is there ever such a thing as truth in the social and cultural world?

Consider this example

"Today is a rainy day."

- Is this a positive, negative or neutral statement?
- It depends on context and perception.
- There is ambiguity in how humans classify things
- We can try to mitigate personal biases by aiming for inter-rater reliability
- But this doesn't mean that we "discovered" the truth
- It merely implies that, we found how much disagreement there is for a group of people in specific places
- Often, models report accuracy scores, measured against ground truth, but how variable is that ground truth itself?

Inter-rater reliability example

- There are many ways to calculate the agreement between human labellers
- Here, we are assuming we have at least 3 labellers and we're going to use Fleiss' κ

```
tp.compare_raters("travel_blogs")
```

✓ 0.0s Python

```
Processing blogs_labeler1...
Processing blogs_labeler2...
Processing blogs_labeler3...

Agreement file saved at travel_blogs_output/agreement.xlsx
Fliess Kappa is 0.6592389670263585
```

See Jupyter Notebook **6.1.ipynb**

Evaluating the results of a model

- Now we are going to see how well an existing model performs against our labelled examples.
- A simple metric is **accuracy**, what percentage of our results correspond to our human-assigned labels?

```
tp.sentiment_analysis("travel_blogs_output/agreement.xlsx")
```

✓ 6.3s Python

```
The accuracy score is 89.0
```

See Jupyter Notebook **6.1.ipynb**

Evaluating the results of a model

- To get a more complete picture, we also need to look at different types of errors, using a confusion matrix
- Let’s look at a simple example, for a classifier that only has two categories: positive and negative

Actual label		Predicted label: positive	Predicted label: negative
	positive	**True Positive**	False Negative
	negative	False Positive	**True Negative**

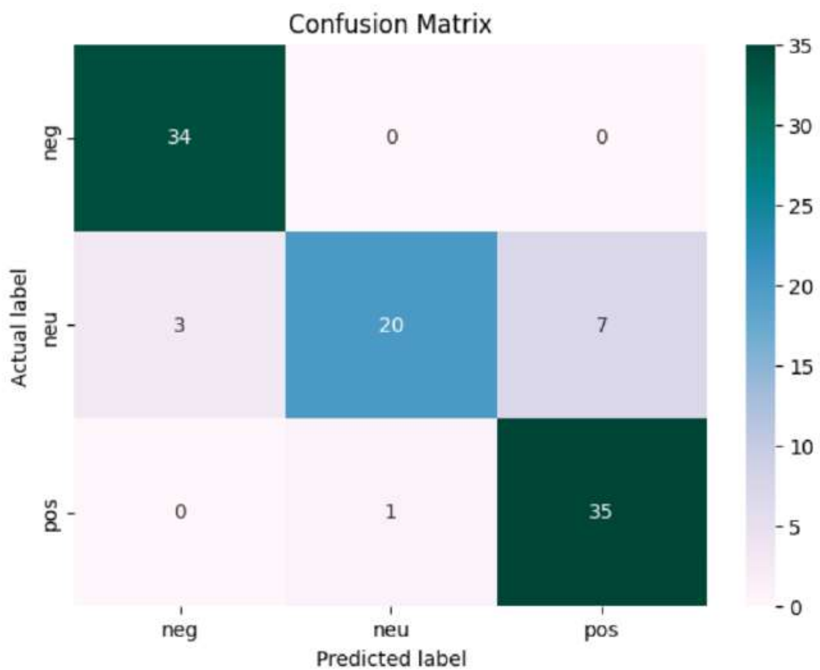
Evaluating the results of a model

- This is a confusion matrix for our earlier sentiment analysis model, which can output three classes: positive, negative and neutral.

```python
sa.confusion_matrix()
```

✓ 0.1s Python

See Jupyter Notebook **6.1.ipynb**



Important note

- In this class, we are not learning how to train a model or how to improve it.
- We are concentrating on the role that interpretation and close-reading play in the design and evaluation of models.
- Why?

Why are we doing this?

- To understand how ‘ground truth’ is always interpretive in the sociocultural world
- Evaluation the output of individual models with our interpretive attention is a crucial skill
- This will be increasingly important in the age of Large Language Models (LLMs)

References

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Paglen, Trevor, and Kate Crawford. n.d. ‘Excavating AI’. -. Accessed 15 September 2023. https://excavating.ai.