SILLM Tutorial 1: LLMs for Content Analysis — Zero and Few-shot Labeling

Indira Sen, David Garcia, Mats Faulborn

25.10.23

Universität Konstanz

Agenda

- Coding environment: Kaggle, Google Colab
- Work with the OpenAI API
- Open-source models (flan-t5)
- Prompting basics
- Label some data with LLMs
    - download datasets from the internet
    - prompt-based labeling / ‘in-context learning’
        - Zero-shot
        - Few-shot
        - Bonus: counterexamples
- Setup OpenAI account [OPTIONAL]
    - If you have a credit card, try out some stuff with the free credits

Why?

- Content analysis:
    - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
    - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990]

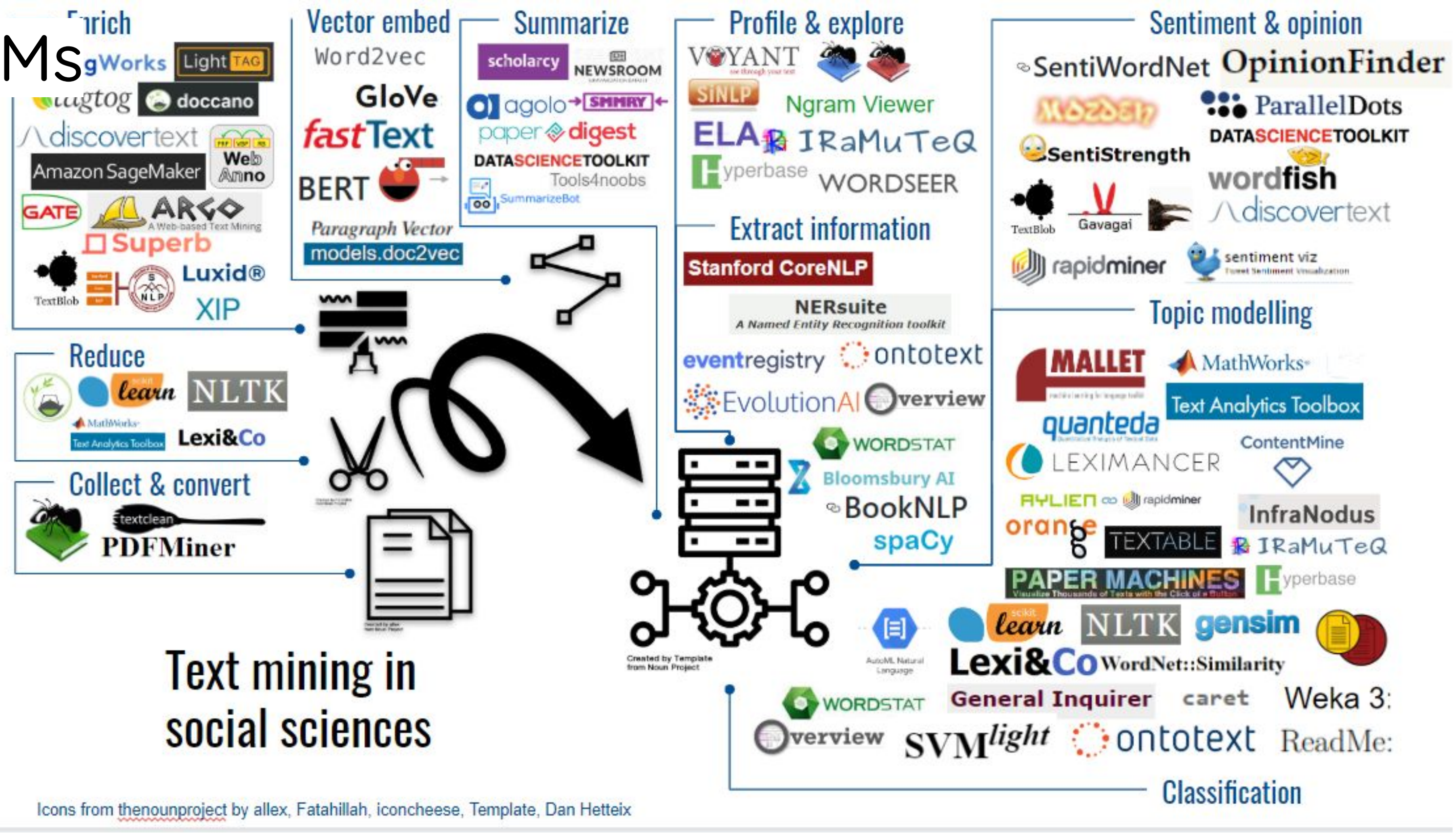
Why?

- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990]
- Questions content analysis can help us answer:
  - How does news media represent the immigration crisis?
  - What are topics that lead to arguments in long-term relationships?
  - How do citizens perceive the performance of politicians during the pandemic?

Why?

- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990]
- Questions content analysis can help us answer:
  - How does news media represent the immigration crisis?
    NYtimes articles frames [Mendelsohn’21]
  - What are topics that lead to arguments in long-term relationships?
    r/relationshipadvice posts topics
  - How do citizens perceive the performance of politicians during the pandemic?
    Twitter posts stance

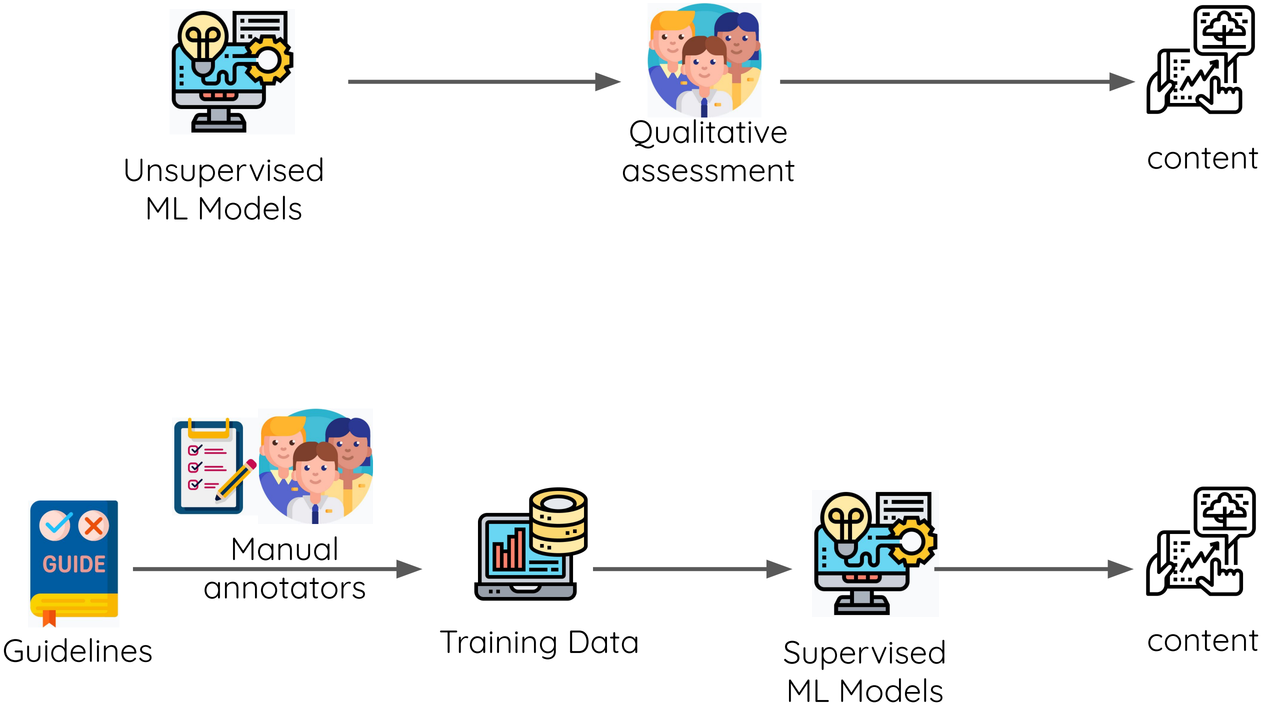
Before LLMs



Taken from PACSS 2020 Sage Ocean Talk

Before LLMs: ‘Modern Pipeline’

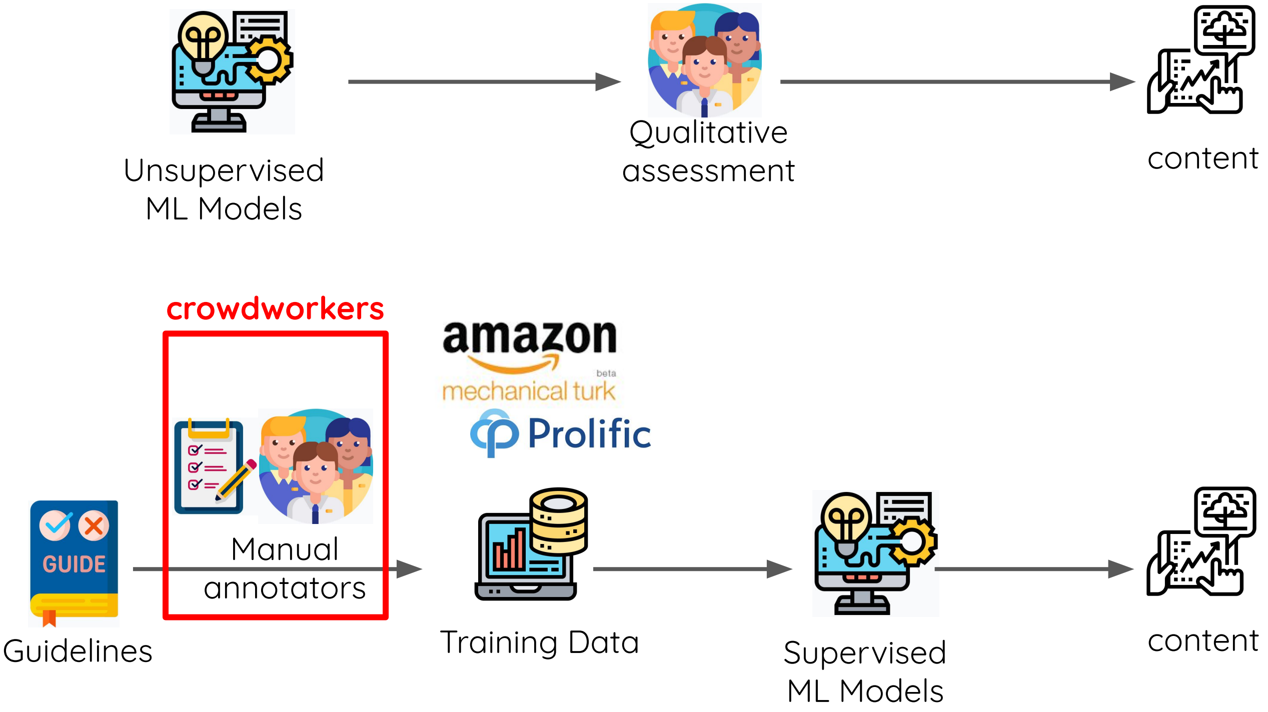
- unsupervised
- supervised



Before LLMs: ‘Modern Pipeline’

- unsupervised

- supervised



After LLMs

Can Large Language Models Transform Computational Social Science?

Caleb Ziems* William Held* Omar Sh
Zhehao Zhang* Diyi Ya
Georgia Institute of Technology, Shanghai Jiao Tong
{cziems, wheld3, jiaaochen}@gatech.edu, zzh12138@sjtu.edu.

BRIEF REPORT | POLITICAL SCIENCES

ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli Authors Info & Affiliations

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

i.org/10.1073/pnas.2305016120

OPEN-SOURCE LARGE LANGUAGE MODELS OUTPERFORM CROWD WORKERS AND APPROACH CHATGPT IN TEXT-ANNOTATION TASKS

After LLMs

Model / Data	Baselines		FLAN-T5					FLAN	Chat	text-001				text-002	text-003
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	ChatGPT	Ada	Babb.	Curie	Dav.	Davinci	Davinci
Utterance Level Tasks															
Dialect	4.5	41.5	1.9	2.3	15.8	16.5	22.6	23.7	15.0	5.3	5.6	6.0	10.9	10.5	16.9
Emotion	16.7	91.7	23.9	65.3	69.1	65.9	66.7	70.3	46.2	44.6	16.1	18.7	19.3	39.8	36.5
Figurative	25.0	94.4	23.6	29.0	25.4	40.2	56.0	64.0	50.2	25.0	24.4	25.0	28.8	52.0	60.6
Humor	50.0	73.1	52.0	51.8	56.2	59.0	50.6	58.8	55.4	55.2	59.0	58.6	50.4	51.4	51.0
Ideology	33.3	61.9	33.1	39.2	48.6	49.2	54.4	48.2	54.8	–	33.3	33.3	34.3	57.6	48.2
Impl. Hate	14.3	69.9	17.7	22.7	17.9	36.3	34.5	35.9	29.7	17.1	18.6	15.7	21.3	22.7	27.1
Misinfo	50.0	82.3	50.0	55.4	69.2	70.2	71.2	77.6	69.0	–	50.4	52.2	52.6	75.6	75.0
Persuasion	12.5	40.4	14.3	19.8	43.9	43.4	†51.6	49.4	40.9	–	16.5	17.0	18.8	26.3	26.3
Sem. Chng.	50.0	65.7	50.3	50.0	†66.9	55.5	51.2	53.7	56.1	50.0	50.5	54.3	39.5	45.9	50.0
Stance	33.3	47.0	34.7	47.8	51.3	52.6	55.9	55.4	†72.0	–	33.1	31.0	48.0	57.4	41.3
Conversation Level Tasks															
Discourse	14.3	47.5	14.7	26.4	37.2	44.3	†52.5	41.9	44.5	13.1	16.5	14.3	17.0	39.8	37.8
Empathy	33.3	33.3	33.3	33.3	35.1	33.7	36.8	†39.8	37.6	–	33.1	35.3	33.3	33.3	33.3
				55.3	†57.1	53.0	53.5	53.2	52.9	50.2	50.0	50.0	50.0	50.8	55.9
				44.2	53.0	59.2	54.2	52.8	50.8	33.1	33.1	32.1	42.2	55.6	47.8
				47.2	50.4	56.8	58.8	60.8	61.6	–	52.2	50.6	49.6	50.5	57.0
				50.6	49.4	54.2	50.0	56.6	53.0	44.6	50.6	49.0	50.8	52.2	51.2
Document Level Tasks															
				–	–	–	-	–	22.3	–	–	8.6	8.6	21.6	22.9
				7.0	1.0	10.9	41.8	50.6	51.3	29.8	47.3	47.4	44.4	48.8	52.4
				34.1	34.1	32.1	49.6	40.3	58.8	32.9	35.1	33.6	25.6	48.7	44.0
Tropes	1.4	0.8	0.9	4.4	8.8	7.9	10.5	16.7	25.4	4.3	7.0	9.6	10.5	18.4	18.4

Table 2: **Zero-shot Classification Results** across our selected CSS benchmark tasks. All tasks are evaluated with accuracy, except for Event Arg. and Event Detection, which use F-1. Models which did not always follow instructions are marked with a dash. Best zero-shot models are in green; zero-shot models that are not significantly worse ($P > .05$; Paired Bootstrap test (Dror et al., 2018)) are marked blue; and † denote cases where zero-shot LLMs match or beat finetuned baselines.

Can Large Language Models Transform Computational Social Science?

Caleb Ziems* William Held* Omar Shaikh* Jiaao Chen*
Zhehao Zhang* Diyi Yang*
Georgia Institute of Technology, Shanghai Jiao Tong University, Stanford University
{cziems, wheld3, jiaaochen}@gatech.edu, zzh12138@sjtu.edu.cn, {oshaikh, diyiy}@stanford.edu

Prompt based labeling

Prompt based labeling

```
[1]  def make_prompt(task, options, instance, **kwargs):
       options_str = '' # options ---> all possible labels
       for i in range(len(options)):
         options_str = options_str + ' %d) %s' %(i+1, options[i])
       prompt = 'Given a piece of text, you have to label whether it is %s or not. Please return one of the following options:%s.' %(task, options_str)


       if kwargs['zero_shot']:
         return prompt + ' What is the label of this text: "' + instance+ '"'
       else: # for few-shot
         examples_str = ''
         for example in kwargs['examples']:
           examples_str = examples_str + 'text: %s, label: %s\n' %(example[0], example[1])
         return prompt + ' Here are some examples of instances and their labels:\n%sWhat is the label of this text: ' %(examples_str) + instance
```

Prompt based labeling w/ OpenAI

- ChatGPT: GPT3.5 Turbo

```
[ ] # ! pip install openai
```

```
[ ] import openai
    openai.api_base="http://91.107.239.71:80" #"http://127.0.0.1:8000"
    openai.api_key="" # enter you API key here

    # list models
    # models = openai.Model.list()
    # models
```

```
responses = openai.ChatCompletion.create(model="gpt-3.5-turbo",
                                         messages=[{"role": "user", "content": prompt}],
                                         max_tokens = 2,
                                         n=runs)
```

Prompt based labeling w/ HuggingFace

- Flan-T5 (small, base, large, XL, *XXL*)

```
[ ] # ! pip install transformers
```

```
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer

model = AutoModelForSeq2SeqLM.from_pretrained("google/flan-t5-xl")
tokenizer = AutoTokenizer.from_pretrained("google/flan-t5-xl", max_new_tokens = 500)
model.cuda()

responses = []
for n in range(0, runs):
    inputs = tokenizer(prompt, return_tensors="pt").to("cuda:0")
    outputs = model.generate(**inputs)
    responses.append(tokenizer.batch_decode(outputs, skip_special_tokens=True)[0])
```

Let’s try it on the notebook:
https://github.com/dgarcia-eu/SILLM/tree/main/Tutorials

Do it yourself

- Now do this for all the instances in your dataset. **Hint**: Use a loop over your dataframe. When doing few-shot labeling, make sure that the examples are not the same as the instance to be labeled.
- Try both zero-shot and few-shot and compare their performance.
- Try both ChatGPT and Flan-T5
- Try to get the label from the LLM output. Is it always as expected and can it always be used as is for quantitative analysis?
- At least for the first 50 instances in your dataset, use metrics like accuracy and F1 score to assess the performance of the LLMs against the true ground truth label.

Bonus:

- try varying the wording of the prompts
- try giving an explicit definition of the task in the prompt