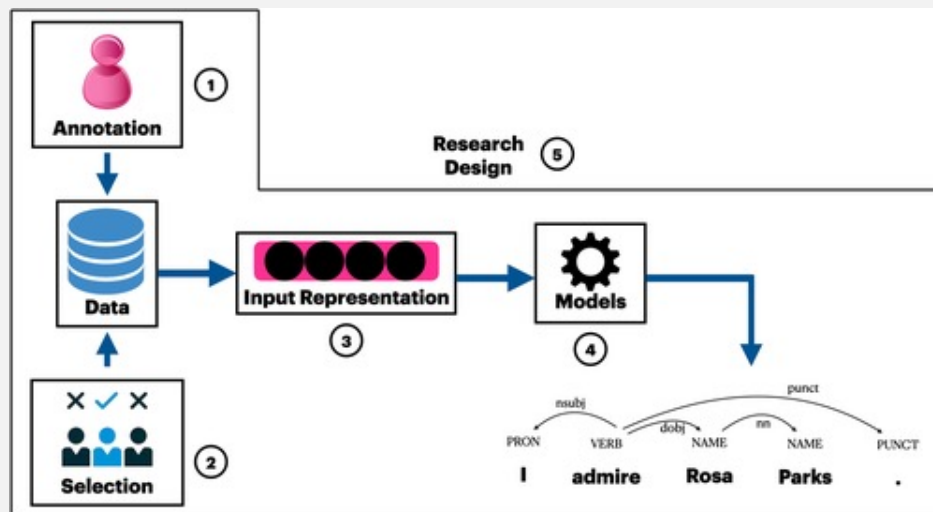


# THE IMPORTANCE OF THE DATASETS IN NLP

Sentiment Analysis Datasets

# IMPORTANCE OF THE DATASETS IN NLP

- sufficient quantity and quality of data and annotations => quality of learned models
- Biases in the data and in the annotation => biases in the models



five sources where bias can occur in NLP systems:

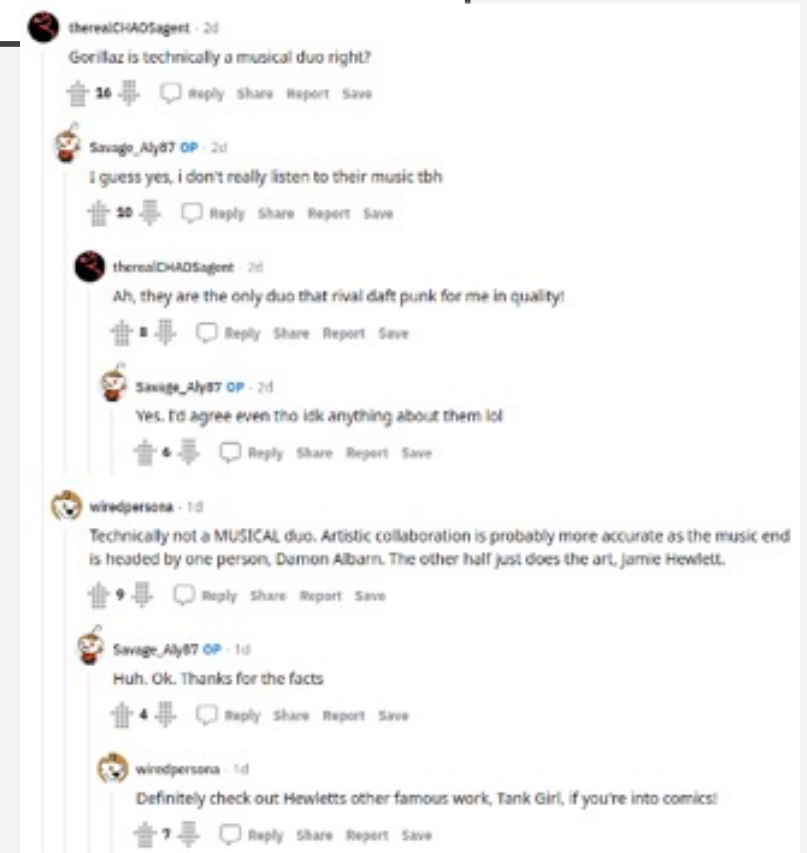
- (1) the data,
- (2) the annotation process,
- (3) the input representations,
- (4) the models
- (5) the research design (or how we conceptualize our research)

From [Hovy and Prabhunoye, 2021]

# EXAMPLES OF PUBLIC DATASETS FOR SENTIMENT ANALYSIS

- For basic sentiment classification
  - Polarity (positive vs. negative)
    - IMDB (see Tutorial/Lab) : movie reviews
  - Emotion categories
    - GoEmotions : written conversations from reddit labelled in 27 emotion categories (annoyance, nervousness, amusement, etc.)

Sample Text	Label(s)
OMG, yep!!! That is the final answer. Thank you so much!	gratitude, approval
I'm not even sure what it is, why do people hate it	confusion
Guilty of doing this tbph	remorse
This caught me off guard for real. I'm actually off my bed laughing	surprise, amusement
I tried to send this to a friend but [NAME] knocked it away.	disappointment



## EXAMPLES OF PUBLIC DATASETS FOR SENTIMENT ANALYSIS

- For Aspect Based Sentiment Analysis
  - product reviews
    - Semeval 2014 Task 4
    - Semeval 2015 Taks 12
    - Semeval 2016 Task 5
  - Twitter data set
    - Dong, Li, et al. "Adaptive recursive neural network for target-dependent twitter sentiment classification."

# EXAMPLES OF PUBLIC DATASETS FOR SENTIMENT ANALYSIS

- Spoken language conversations
  - [SILICONE](#) benchmark

## **SILICONE BENCHMARK**

The Sequence labelling evaluation benchmark for spoken language (SILICONE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems specifically designed for spoken language. All datasets are in the English language and cover a variety of domains including daily life, scripted scenarios, joint task completion, phone call conversations, and television dialogues. Some datasets additionally include emotion and/or sentiment labels.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, Chloe Clavel, [Hierarchical Pre-training for Sequence Labelling in Spoken Dialog](#), Findings of EMNLP 2020,

## EXAMPLES OF PUBLIC DATASETS FOR SENTIMENT ANALYSIS

- Hate speech

Topic	Dataset	Abbrev.	Total Size	Pos. Size	Ratio
Topic-Generic	Davidson	Gene Davi	5590	1430	25.58%
	Founta	Gene Fnt	57355	4119	7.18%
Topic-Specific (Gender)	Evalita	Gndr Evit	5000	2245	44.90%
	HatEval <sub>women</sub>	Gndr HatE	6472	2845	43.96%
	IberEval	Gndr Iber	3977	1851	46.54%
	Waseem <sub>sexism</sub>	Gndr Wasm	14531	3216	22.13%
Topic-Specific (Race)	HatEval <sub>immigrants</sub>	Race HatE	6499	2617	40.27%
	Waseem <sub>racism</sub>	Race Wasm	13272	1957	14.75%

*[What Did You Learn To Hate? A Topic-Oriented Analysis of Generalization in Hate Speech Detection](<https://aclanthology.org/2023.eacl-main.254>) (Bourgeade et al., EACL 2023)*

# HOW TO BUILD A LABELLED DATASET?

- The two steps
  - Data collection
  - Data annotation

# DATA COLLECTION

First important question:

what type of data is being used?

=> what linguistic phenomena are present in the data used to train your models?

- spoken (using transcripts)/written
- w/o interactions (monologues, dialogues)
- Natural/collected through a real application vs. simulated/scripted/prepared

In real applications, corpora contain spontaneous expression and can be 'wild' [Schuller et al., 2016] (i.e. contain noisy text)

- Ex1: Spoken transcripts of call-centre data contain disfluencies
- Ex 2: written conversations contain typos, or chat features: lol, A +, mouhahaha

Disfluences combinées Vous regardez les 5 derniers chiffres des  
chi des numéros gravés, pas les chiffres qui défilent hein

lol, A +, mouhahaha

« I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, that all men are created equal." » ....

A: I'm worried about something.  
B: What's that?  
A: Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.  
B: That's annoying, but nothing to worry about. *Just breathe deeply when you feel yourself getting upset.*  
A: Ok, I'll try that.  
B: Is there anything else bothering you?  
A: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.  
B: Do you have any other plans this weekend?  
A: I'm supposed to work on a paper that'd due on Monday.  
B: *Try not to take on more than you can handle.*  
A: You're right. I probably should just work on my paper. Thanks!

Figure 1: An example in **DailyDialog** dataset. Some text is shortened for space. Best viewed in color.



# DATA COLLECTION

Second question: what protocol for data collection?

- Data **collected via crowdsourcing platforms** (workers)
  - Ex: Empathetic Dialogues (Rashkin et al., 2018)
    - Some workers are asked to start the conversation following an emotional prompt.
    - Others have to reply.
- Open data **available on the web** and on social networks
  - Ex GoEmotions retrieved from reddit
- Data **collected within companies**
  - Ex : call-center transcripts, chatbot interactions, complaint emails

# HOW TO BUILD A LABELLED DATASET?

- The two steps
  - Data collection
  - Data annotation :
    - To supervise the training machine learning models
    - To evaluate the models
      - Fine-grained annotation can also provide a better understanding of the behavior of the ML models by accessing to finer content

# DATA ANNOTATION EXAMPLE

TABLE I  
EXAMPLES OF TWEET-TARGET PAIRS FROM SEMEVAL2016-T6 DATASET AND COVID-19-STANCE DATASET.

Dataset	Tweet	Target	Stance	Emotion
SemEval2016-T6	Job should always go to best candidate, regardless of gender. Gender shouldn't even matter anymore, it's 2015! #PaulHenry #SemST	Feminist Movement	In Favor	Positive
SemEval2016-T6	We are actually watching a video on radical feminism in history this is the funniest movie ive ever seen. #SemST	Feminist Movement	Against	Positive
COVID-19-STANCE	@realDonaldTrump @Mike_Pence What a disaster of a group. Not everyone is wearing masks. What are you people thinking?! Lead by example.	Wearing a Face Mask	In Favor	Negative
COVID-19-STANCE	It's amazing how many people just roll over and wear masks despite a preponderance of evidence that they dont help nor are they even necessary.	Wearing a Face Mask	Against	Negative

To define:

- The annotation unit (sentence, document, word)

- The labels

# HOW TO CHOOSE SENTIMENT LABELS?

$y_i \in ???$

- Positive/negative classes are too vague:
  - Negative emotion? « I am sad»
  - Negative opinion? « I don't like this plot »
  - Negative mood? « I am in a bad mood, today»



**Emotion, mood, opinion, are different phenomena...**

Clavel, C.; Callejas, Z., [Sentiment analysis: from opinion mining to human-agent interaction](#), *Affective Computing, IEEE Transactions on*, 7.1 (2016) 74-93.

# HOW TO CHOOSE SENTIMENT LABELS?

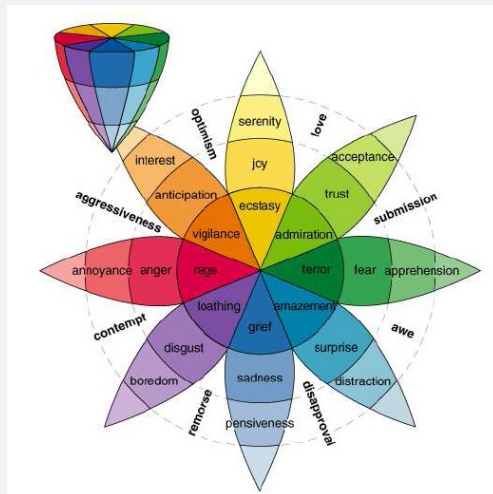
## Categories vs. dimensions

- Use sentiment categories (annoyance, frustration, etc.)
- OR use dimensions
  - Valence (ex: -3,-2,-1,0,1,2,3)
  - Intensity (ex: 0,1,2,3)
- OR combine both of them

Relevant classes depend on the **applications** and on the **studied data**

## RELEVANT OPINION/EMOTION PHENOMENA DEPEND ON THE APPLICATIONS AND STUDIED DATA

- Ex: Emotions (short phenomenon, physiological reaction, appraisal of a major event (stimulus)) are used for example for:
  - Call-center analytics: Anger, dissatisfaction



Model frequently used for emotions: choose classes based on the cone of Plutchik's emotions [Plutchik, 1984].



- E-learning: Frustration, boredom...



## RELEVANT OPINION/EMOTION PHENOMENA DEPEND ON THE APPLICATIONS AND STUDIED DATA

- Ex: Mood (diffuse non-caused low-intensity long-duration change in subjective feeling) is used for companion robot at home (Detection of Gloomy/depressed states)



## RELEVANT OPINION/EMOTION PHENOMENA DEPEND ON THE APPLICATIONS AND STUDIED DATA

- Ex: Stance (in favor or against a specific target) is used for opinion analysis in social networks

TABLE I  
EXAMPLES OF TWEET-TARGET PAIRS FROM SEMEVAL2016-T6 DATASET AND COVID-19-STANCE DATASET.

Dataset	Tweet	Target	Stance	Emotion
SemEval2016-T6	Job should always go to best candidate, regardless of gender. Gender shouldn't even matter anymore, it's 2015! #PaulHenry #SemST	Feminist Movement	In Favor	Positive
SemEval2016-T6	We are actually watching a video on radical feminism in history this is the funniest movie ive ever seen. #SemST	Feminist Movement	Against	Positive
COVID-19-STANCE	@realDonaldTrump @Mike_Pence What a disaster of a group. Not everyone is wearing masks. What are you people thinking?! Lead by example.	Wearing a Face Mask	In Favor	Negative
COVID-19-STANCE	It's amazing how many people just roll over and wear masks despite a preponderance of evidence that they dont help nor are they even necessary.	Wearing a Face Mask	Against	Negative



## CHOICE OF THE ANNOTATION UNIT

- Depending on the task and on the data
  - Word
  - Chunk
  - Sentence
  - Inter-pausal unit for speech transcript

## DEFINE THE CONTEXT TO BE PROVIDED TO THE LABELLER

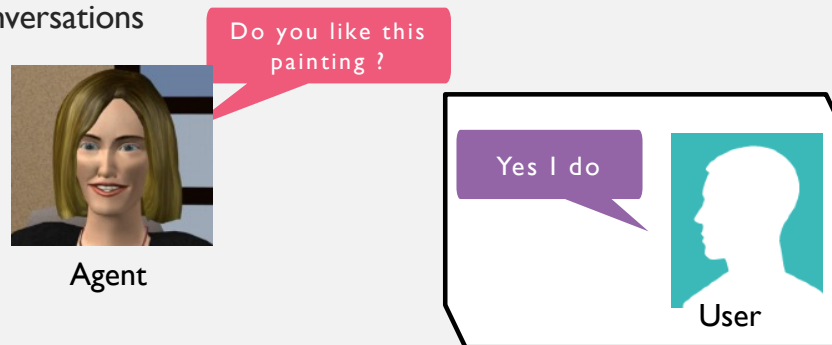
Examples of the importance of context in which the utterances were expressed

- Societal context in twitter :
  - who is the writer?
  - what is its claimed political affiliation?
- ⇒ help to better understand the opinion expressed
- ⇒ Annotation in the context or keeping the contextual information in meta-data

## DEFINE THE CONTEXT TO BE PROVIDED TO THE LABELLER

Examples of the importance of context in which the utterances were expressed

- Interaction context in conversations



- ⇒ help to better understand the opinion expressed (in this example, difficult to annotate without context)
- ⇒ Annotation in the context

# DEFINE THE CONTEXT TO BE PROVIDED TO THE LABELLER

=> Annotation in the context

Previous Annotations

1

AGENT

hi

USER

how are you today

LIKES

DISLIKES

No expression of like

No expression of dislike

UTTERANCES 2

AGENT

how's it going?

USER

it's going great

LIKES

DISLIKES

No expression of like

No expression of dislike

The Annotation Task

2

Current Utterances 3/20

AGENT

What is your name?


USER

my name is alex what is your name

3


QUESTION 1: IN THIS DIALOGUE, DOES THE USER EXPRESS THAT HE/SHE LIKES ONE OR SEVERAL OBJECTS ITEMATIZED BELOW?

☐




The Clock

☐




One Set of Plates

☐




The Both Sets of Plates


☐



☐



☐



# DEFINE THE CONTEXT TO BE PROVIDED TO THE LABELLER

Examples of the importance of context in which the utterances were expressed



- Multimodal context influence the perception of the sentiment
  - Verbal content
  - Audio (prosody, voice quality)
  - Video (gesture, posture, facial expressions)

⇒ help to better understand the opinion expressed

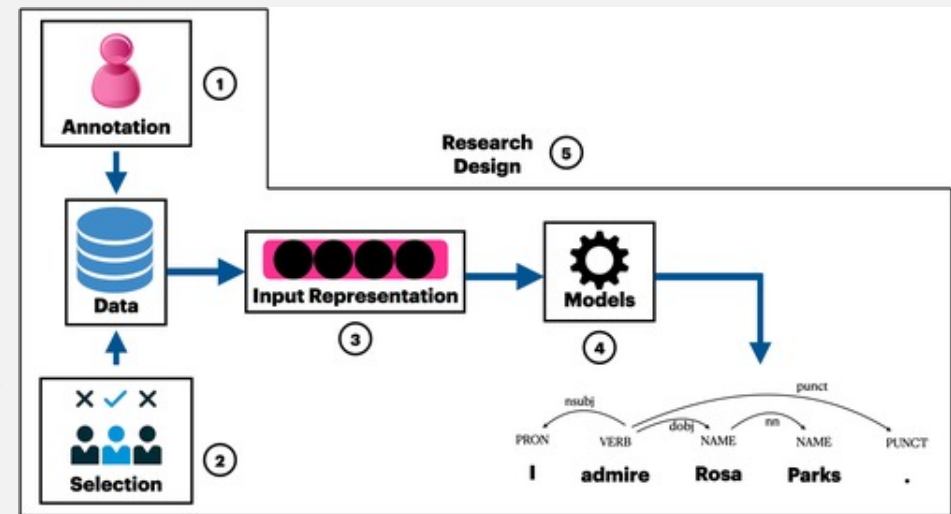
⇒ Annotation in the multimodal context



# STORE SOCIAL VARIABLES OF LABELLERS

It is important to understand the annotation biases that will be encoded in the model => Store the social variables of labellers

- Personality questionnaires
  - (Big Five Model): Openness Conscientiousness Extraversion Agreeableness Neuroticism
- Socio-demographic criteria : age, gender, education
- EXAMPLE : POM dataset [Park et al., 2014] self-assessed personality of the workers (labellers)



# MEASURE THE RELIABILITY OF ANNOTATIONS

Opinion/Emotion phenomenon = subjective phenomenon

- We do not all have the same perception of an opinion expressed by the other
- TODO:
  - Multiple annotators
  - Assess the degree of reliability of the annotations

# MEASURE THE RELIABILITY OF ANNOTATIONS

## Measures

- Cohen's kappa [Carletta, 1996]:
- agreement corrected for what it would be under the mere fact of chance

$$k = \frac{p_o - p_e}{1 - p_e}$$

- $p_o$  is the proportion of agreement observed and  $p_e$  the probability that the annotators agree by chance



## MEASURE THE RELIABILITY OF ANNOTATIONS

EXO

- Kappa values ?
  - When annotators agree as much as chance
  - When the annotators agree totally
- Exercice
  - 50 text sequences annotated by 2 people (Ann1 / Ann2) in 2 categories positive / negative
  - Calculating kappa between the two annotators

$$k = \frac{p_o - p_e}{1 - p_e}$$

Ann1 \ Ann2	Positive	Negative
Positive	20	5
Negative	10	15

## MEASURE THE RELIABILITY OF ANNOTATIONS

- $P_o = (20+15)/50 = 0,7$
- Calculate  $P_e$ :
  - Ann1 uses positive label 50% of the time
  - Ann2 uses positive label 60% of the time
  - Probability that Ann1 and Ann2 use the positive label:  $0.5*0.6=0.3$
  - Probability that Ann1 and Ann2 use the negative label :  $0.5*0.4 = 0.2$
  - Probability to agree by chance :  $0.2+0.3 = 0.5$
- Kappa computation:
  - $Kappa = 0.2/0.5 = 0.4$

$$k = \frac{p_o - p_e}{1 - p_e}$$

## MEASURE THE RELIABILITY OF ANNOTATIONS

- Moderate agreement = standard for emotions [Landis & Koch, 1977]
- Other measures:
  - Fleiss Kappa : variant with random raters
  - ICC : Intraclass correlation coefficient
    - Useful existing variant with random raters (frequent when using crowdsourcing platforms)
  - Cronbach's Alpha [Cronbach, 1951] for dimensions

Kappa value	Interpretations
<0	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

To go further : Computing Inter-Rater Reliability for Observational Data:An Overview and Tutorial Kevin A. Hallgren

## SEMI-AUTOMATIC ANNOTATION

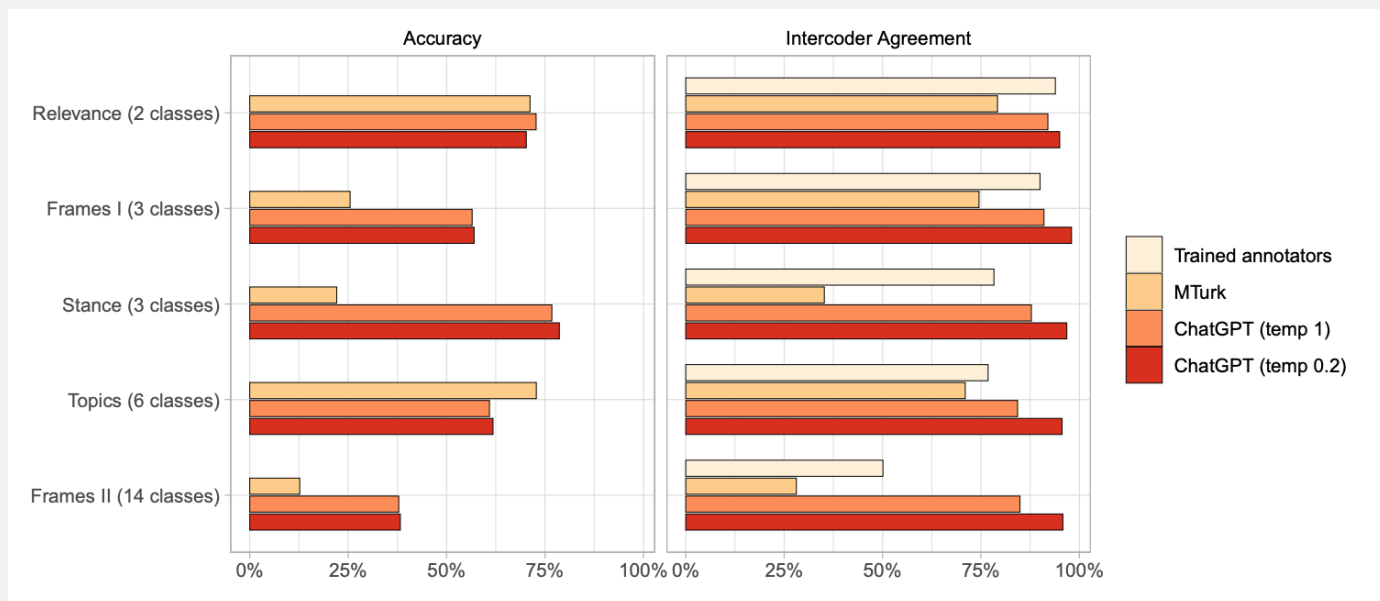
- Objective: lighten the human annotation workload and reduce annotation cost
- Approach: Combine manual annotation with automatic annotation
  - Ex: to prevent having the human judges select among 42 emotional categories use a classifier trained on close datasets to output the 3-best labels

A Large-Scale Dataset for Empathetic Response (Welivita et al., EMNLP 2021)

And very recently...

## AUTOMATIC ANNOTATION WITH CHATGPT

- « ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks »  
<https://arxiv.org/pdf/2303.15056.pdf>



Temperature in order to control the creativity level :  
Level close to 0 : more conservative and predictable outputs  
Level close to 1 : more diverse and unpredictable outputs

# DATA ANNOTATION

- A few tips
  - Build an annotation guide
    - Go beyond the positive/negative classification and define among the wide variety of existing Opinion/emotion types...
    - ... the types and the categories that will be relevant for your application (anger OR discontent)
    - Use of questions to guide the annotation process
    - Give examples of categories
  - Use sufficiently-expert annotator on crowdsourcing platforms
    - Ex. of annotation tools : gradio, Gate, ELAN, Webanno (sometimes you will have to develop your own tool)
    - Ex. of crowdsourcing platform : prolific.ac
  - Measure the reliability of annotations
  - If you use automatic annotation: check the performance of the automatic by comparing to human annotation performance !