language and translation technology team

GHENT UNIVERSITY

Handling disagreement

An introduction to perspectivism

Before model training: Crafting high-quality text annotations for ML

Presentation by Sofie Labat
June 3rd, 2024
Tutorial at ICWSM'24

Outline

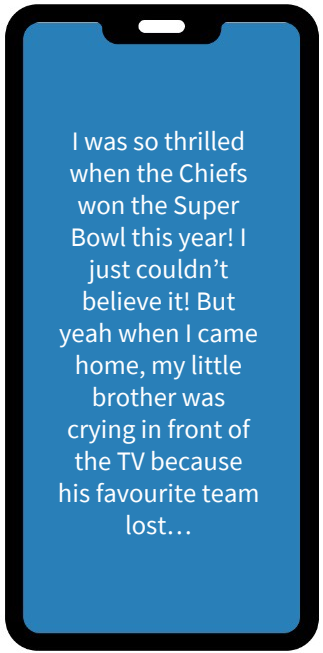
How do we interpret and handle low inter-annotator agreement?

1. Questionnaire design
2. Errors by annotators
3. Human label variation and perspectivism

#1 Questionnaire design

Question phrasing

Example (1)

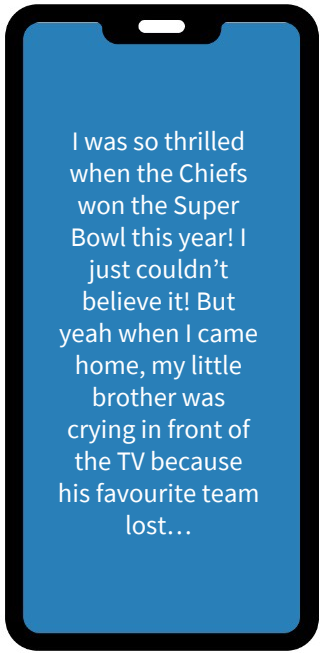


Task 1

What emotion expresses the text?

Question phrasing

Example (1)



Task 1

What emotion expresses the text?

Task 2

What emotion feels the writer? Joy, Excitement

Task 3

What emotion does his brother feel? Sadness, Disappointment

Task 4

What emotion do you as writer feel when reading the text? Neutral

Question phrasing

Example (2)

Task 1

Is the label CANOE in the image?

Participant responses (%)		
Yes	Unsure	No
15.8	10.5	**73.7**

Source: “Is a picture of a bird a bird? A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models”. Parrish et al. (2024).

Question phrasing

Example (2)

Task 1

Is the CANOE in the image?

Participant responses (%)		
Yes	Unsure	No
15.8	10.5	**73.7**

Task 2

The machine learning model predicted CANOE.
Is the model correct?

Participant responses (%)		
Yes	Unsure	No
71.4	7.1	21.4

Source: “Is a picture of a bird a bird? A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models”. Parrish et al. (2024).

Question phrasing

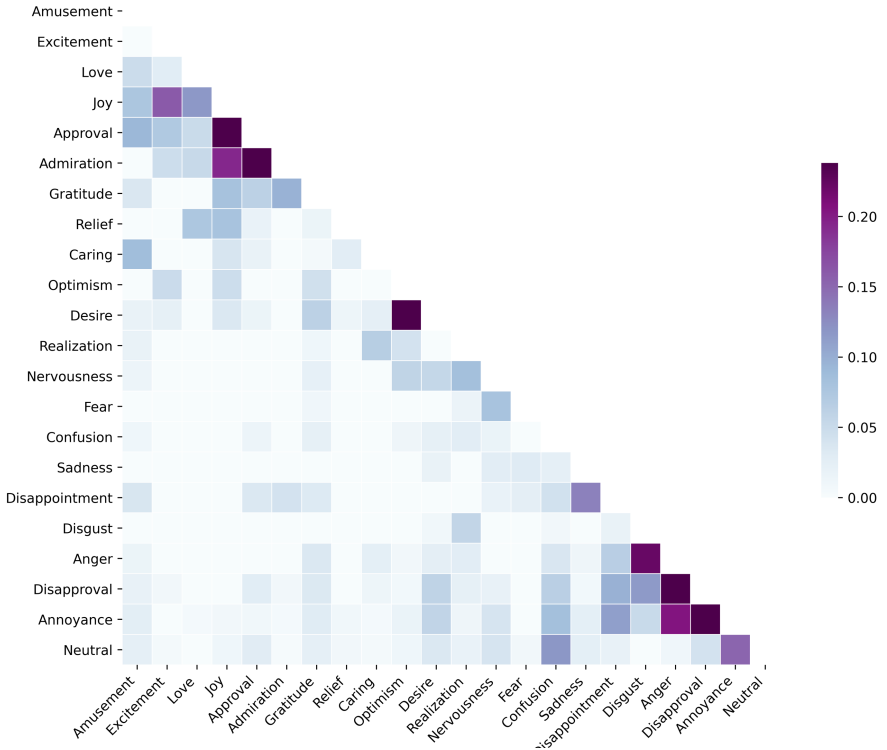
Guidelines

- **Avoid ambiguity**
  See previous examples
- **Be specific & concise**
  Long texts will less likely be read by annotators
- **Repeat key information**
  Repetition increases information absorption
- **Explicitly state rejection criteria**
  Improvement of data quality & less data cleaning
- **Writing style**
  Avoid passives, complex sentence structures, difficult terminology...

Choice of labels

Potential bottlenecks

- **Imbalance**
  Imbalance is common in real-world data
- **Semantic overlap**
  Certain labels have overlapping meaning
- **Domain-dependent taxonomies**
  Differing distributions depending on the domain
- **Theory does not always uphold in practice**
  Do not just apply theoretical taxonomies on real-world data

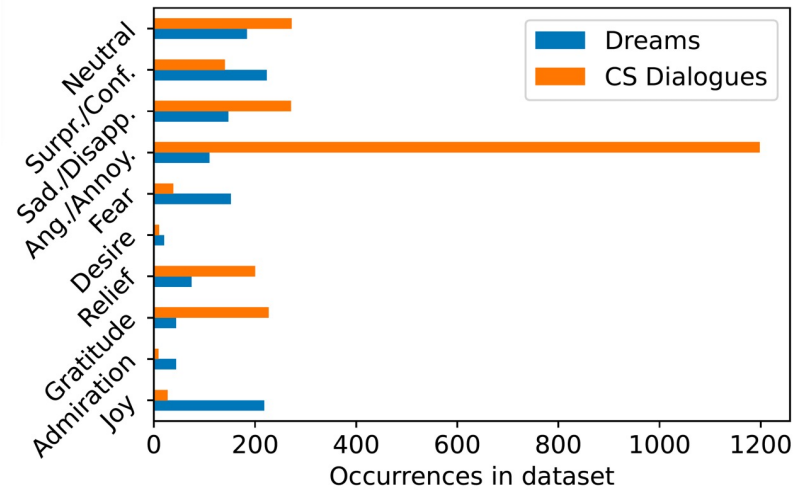


Source: “EmoTwiCS: A Corpus for Modelling Emotion Trajectories in Dutch Customer Service Dialogues on Twitter”. Labat et al. (2023).

Choice of labels

Potential bottlenecks

- **Imbalance**
  Imbalance is common in real-world data
- **Semantic overlap**
  Certain labels have overlapping meaning
- **Domain-dependent taxonomies**
  Differing distributions depending on the domain
- **Theory does not always uphold in practice**
  Do not just apply theoretical taxonomies on real-world data



Source: “EmoProgress: Cumulated Emotion Progression Analysis in Dreams and Customer Service Dialogues”. Wemmer et al. (2024).

Choice of labels

Strategies to address these bottlenecks in the questionnaire

- **Clarify labels**
  Add definitions that distinguish between labels
- **Hands-on illustrations**
  Give examples in the annotation scheme
- **Cluster labels**
  If certain labels are often confused with one other, make a joined label
- **Update taxonomy iteratively**
  Run prestudies to test the suitability of the annotation labels

#2 Errors by annotators

Errors by annotators

How to reduce human errors & low effort submission in the annotation process?

- **Text annotation (closed-ended)**
  Include attention checks
- **Text generation (open-ended)**
  Provide an overview of explicit criteria for rejection
  Prohibit use of ChatGPT (see Veselovsky et al. (2023))
- **Assess annotator skills**
  Preliminary tests to screen annotators for specific task (e.g., assess one's emotional intelligence in practical task)

Source: “Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks”. Veselovsky et al. (2023).

#3 Human label variation and perspectivism

Human label variation

Every humanly labeled dataset contains variation

Inter-annotator disagreement

Certain human attributes correlate with variation between annotators

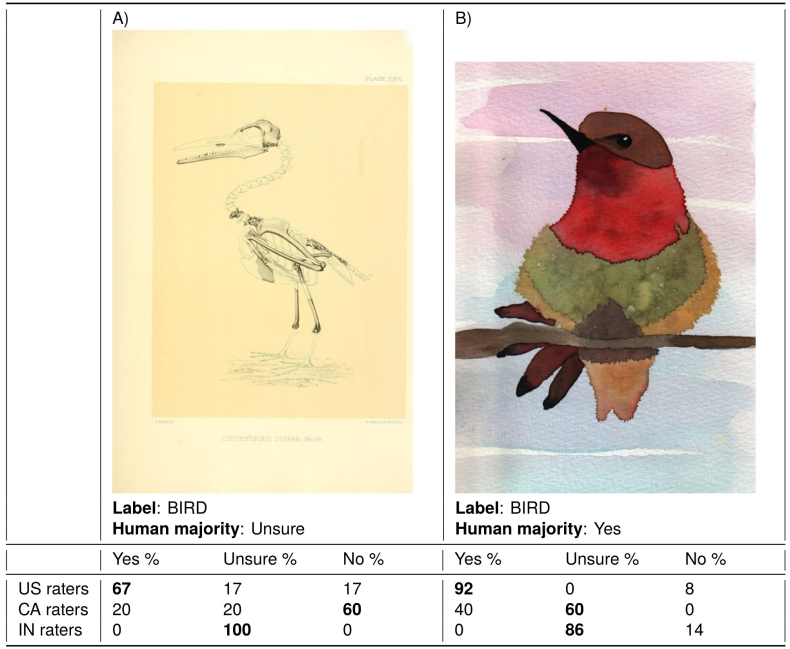
Intra-annotator disagreement

Our own opinions diverge over time

Inter-annotator disagreement

Which human characteristics influence inter-annotator disagreement?

- **Geographical locale**
  Locales can influence label choices
- **Personality**
  Annotators who are extraverted and conscientious are better at labelling emotions
- **Beliefs and identities**
  Bias in toxic language detection
- **Many other socio-demographic variables...**



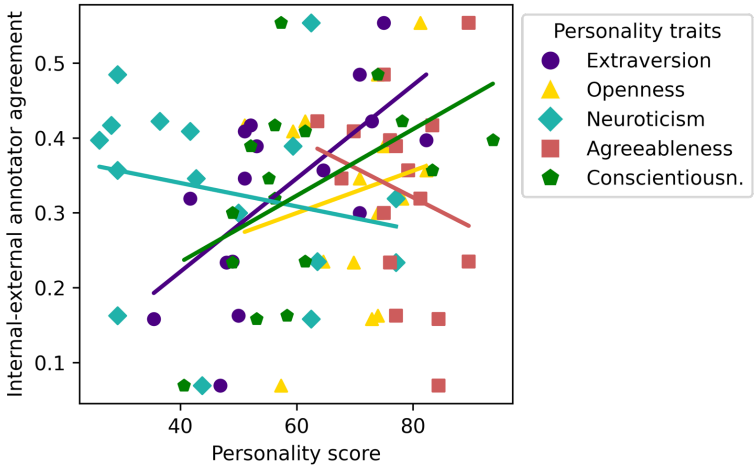
	A)			B)		
	Label: BIRD **Human majority**: Unsure			**Label**: BIRD **Human majority**: Yes		
	Yes %	Unsure %	No %	Yes %	Unsure %	No %
US raters	**67**	17	17	**92**	0	8
CA raters	20	20	**60**	40	**60**	0
IN raters	0	**100**	0	0	**86**	14

Source: “Is a picture of a bird a bird? A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models”. Parrish et al. (2024).

Inter-annotator disagreement

Which human characteristics influence inter-annotator disagreement?

- **Geographical locale**
  Locales can influence label choices
- **Personality**
  Annotators who are extraverted and conscientious are better at labelling emotions
- **Beliefs and identities**
  Bias in toxic language detection
- **Many other socio-demographic variables...**



Source: “Variation in the Expression and Annotation of Emotions: a Wizard of Oz Pilot Study”. Labat et al. (2022).

Inter-annotator disagreement

Which human characteristics influence inter-annotator disagreement?

- **Geographical locale**
  Locales can influence label choices
- **Personality**
  Annotators who are extraverted and conscientious are better at labelling emotions
- **Beliefs and identities**
  Bias in toxic language detection
- **Many other socio-demographic variables...**

Anti-Black posts		*Rated as Offensive*	*Rated as Racist*
EMPATHY		$r =$ 0.285 **	$r =$ 0.286 **
ALTRUISM		$r =$ 0.380 **	$r =$ 0.441 **
HARMOFHATESPEECH		$r =$ 0.451 **	$r =$ 0.528 **
FREEOFFSPEECH		$r =$ -0.394 **	$r =$ -0.467 **
RACISTBELIEFS		$r =$ -0.513 **	$r =$ -0.574 **
LINGPURISM		$r =$ -0.154 **	$r =$ -0.167 **
TRADITIONALISM		$r =$ -0.206 **	$r =$ -0.237 **
Politics	*(lib.: 0, cons.: 1)*	$r =$ -0.374 **	$r =$ -0.441 **
Gender	*(men: 0, women: 1)*	$d =$ 0.321 **	$d =$ 0.341 **
Race	*(White: 0, Black: 1)*	$d =$ 0.301 *	*n.s.*

Table 3: Associations between annotator variables and ratings of offensiveness and racism for the *anti-Black* posts in the *breadth-of-workers* study. We use the Holm correction for multiple comparisons for non-hypothesized associations and only present significant Pearson r or Cohen's d effect sizes (*: $p < 0.05$, **: $p < 0.001$; *n.s.*: not significant).

Source: "Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection". Sap et al. (2022).

Intra-annotator disagreement

Do we vary in our own annotations over time?

- **Intra-rater agreement deteriorates over time**

  After two weeks, this deterioration stabilizes (finding also supported in previous work from Kiritchenko & Mohammad (2017), Li et al. (2010))

	Billingual participants			**Majority vote *v.* Original labels**	
	Fleiss	Cohen	%	Cohen	%
All	0.28	0.29	64.2	0.44	71.7
EN	0.29	0.29	64.6	0.48	74.0
DE	0.27	0.27	63.4	0.40	68.9

Table 1: Reliability as measured by inter-annotator agreement (Fleiss' and Cohen's κ and raw percentage agreement). Cohen's κ and % are calculated pairwise.

		κ	%
All items		0.49	74.5
Same language	All	0.44	72.3
	EN	0.43	71.6
	DE	0.45	72.9
Different language	All	0.53	76.9
	EN→DE	0.54	77.2
	DE→EN	0.53	76.6

Table 2: Stability as measured by intra-annotator agreement (Cohen's κ and raw percentage agreement).

Source: "Temporal and Second Language Influence on Intra-Annotator Agreement and Stability in Hate Speech Labelling". Abercrombie et al. (2023).

Multiple perspectives in ML

Methods to include variation in human annotations at ML

- **Various models for the same task**
  One model per group of personality traits/characteristics (Casola et al., 2023)
- **Include profiling information in the model**
  Profiling information can be passed to the model (Wan et al., 2023)
- **Human calibration error**
  Evaluate when a classifier knows when it does not know, in the context of human variation (Baan et al., 2022)

Sources: “Confidence-based Ensembling of Perspective-aware Models”. Casola et al. (2022).
“Everyone’s Voice Matters: Quantifying Annotation Disagreement Using Demographic Information”. Wan et al. (2023).
“Stop Measuring Calibration When Humans Disagree”. Baan et al. (2022).

Questions?