University of Bamberg

Appraisals as Emotion Model in NLP

Why we need them and how to acquire data

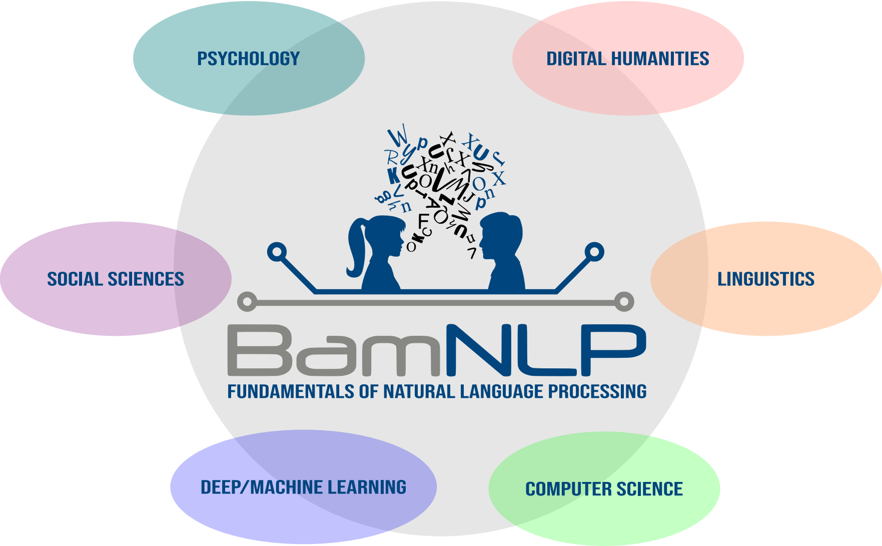
Workshop, Ghent, September 26, 2025

Roman Klinger
roman.klinger@uni-bamberg.de

romanklinger.de romanklinger
https://www.bamberg.de/nlproc/

About Myself

- 1999–2006: Studies at University of Dortmund:
  Computer science with minor psychology
- 2006–2010: Doctoral studies at Fraunhofer SCAI, St. Augustin:
  Biomedical text mining, machine learning
- 2010, 2013: Research visits at UMass Amherst:
  Probabilistic machine learning, MCMC inference
- 2011–2012: Postdoc at Fraunhofer SCAI:
  Social media mining, eGovernment
- 2013–2014: Postdoc at Bielefeld University:
  Sentiment analysis, opinion mining
- 2015: Co-Founder of Semalytix GmbH (exit 2020)
  Social Media Health Mining
- 2014–2024: (Senior) Lecturer/apl. Prof at IMS, Uni Stuttgart
  Natural Language Understanding and Generation
- 03/2024: Full Professor for Fundamentals of NLP, Bamberg

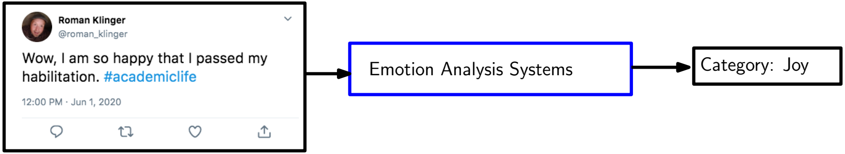

PSYCHOLOGY
DIGITAL HUMANITIES
SOCIAL SCIENCES
LINGUISTICS
BamNLP
FUNDAMENTALS OF NATURAL LANGUAGE PROCESSING
DEEP/MACHINE LEARNING
COMPUTER SCIENCE


Digitales Lehren & Lernen
KI & Gesundheit
KI & Ethik
KI & Bildung
Digitale Überlieferung
Kognition & Interaktion
Soziotechnische Systeme
Bamberger Zentrum für Künstliche Intelligenz
Erklärbares Maschinelles Lernen
KI-System-entwicklung
Kognitive Systeme
Kultur-informatik
Medien-informatik & Information Retrieval
Sprach-generierung & Dialogsys-teme
Smart Environ-ments
User Experience & Design
Informa-tionsvisua-lisierung
Grundlagen der Sprach-verarbeitung
Multimodale Intelligente Interaktion
Mensch-Computer-Interaktion
Algorithmen & Komplexi-tätstheorie
KI-Enginee-ring in Unter-nehmen
Computa-tional Social Science & KI
Computer-grafik & ihre Grundlagen
Data Engineering
Energie-effiziente Systeme
Analyse sozialer Netzwerke

Outline

Outline

Emotion Analysis: What we want to do.



Emotion Examples

Which emotion was felt by the author of the examples?

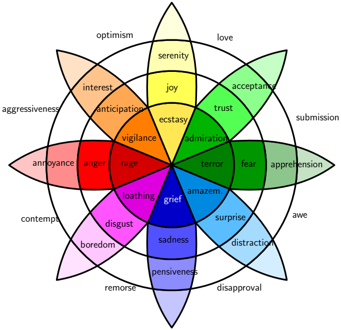
How did you recognize that?

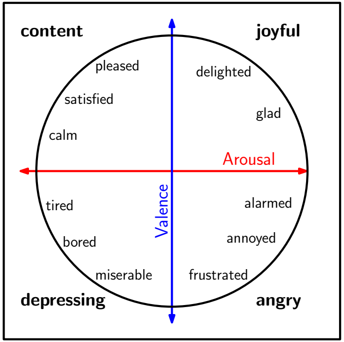
- “She became angry.”
- “A tear was running down my face.”
- “Their dog ran towards me quickly.”

With this exercise, we discussed:

- What is an appropriate set of emotions?
- How are they expressed/recognized?
- Emotions are subjective.

How to define a categorical system of emotions?

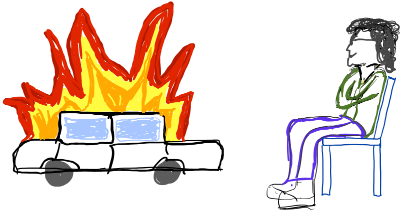




Definition of Emotions: Components

Emotion (Scherer, 2005)

Emotions are "an episode of interrelated, synchronized changes in the states of [...] five organismic subsystems in response to the evaluation of a [...] stimulus-event ..."



Event

Feeling Expression Bodily Symptom

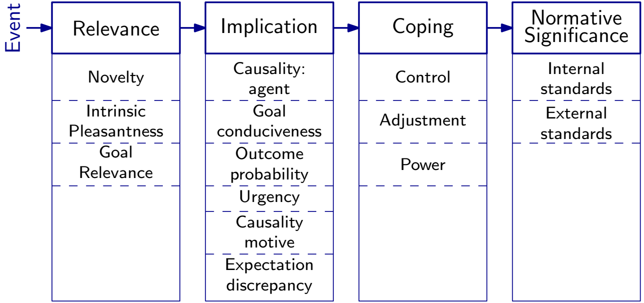
Action Tendency Cognitive Appraisal

Components

Fear

Name

Cognitive Appraisal in Scherer's Component Process model



K.R. Scherer (2001). Appraisal Considered as a Process of Multilevel Sequential Checking.

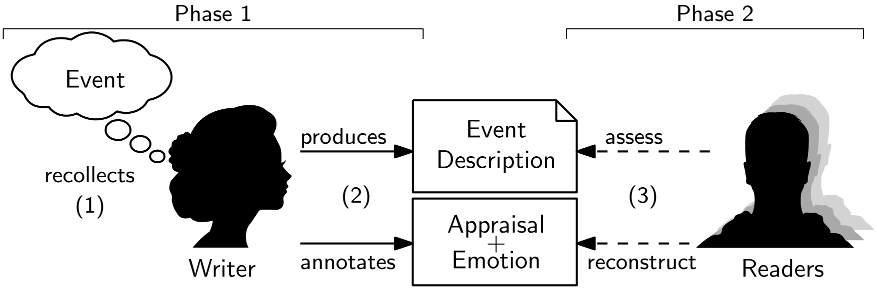
Research Questions

- Can appraisals and emotions be annotated reliably by external annotators?
- Can we computationally model appraisals and does it help emotion categorization?

E. Troiano et al. (2023). "Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction". In: Computational Linguistics 49.1

J. Hofmann et al. (2020). "Appraisal Theories for Emotion Classification in Text". In: COLING

Approach



- Production: 550 event descriptions for anger, boredom, disgust, fear, guilt/shame, joy, pride, relief, sadness, surprise, trust, no emotion
- Five readers for subset of produced texts

Examples

pride I baked a delicious strawberry cobbler.

fear I felt ... when there was a power outage in my home. That day, my wife and I were cuddling in the sitting room when a thunderstorm started. Then ... filled me when thunder hit our roof and all the lights went off.

joy I found the perfect man for me, and the more time goes on, the more I realized he was the best person for me. Every day is a ....

Reliability Results

		Agreement						
			Emotion F$_1$		Emotion Acc.		Appraisal RMSE	
Condition	Val.	#Pairs	G–V	V–V	G–V	V–V	G–V	V–V
All Data		6600 12000	.49	.50	*.49	*.52	*1.57	*1.48
Gender match	M–M	631 1113	.50	*.45	.51	*.49	1.55	1.50
	F–F	2405 1377	.49	*.52	.51	*.55	1.57	*.1.50
	≠	2962 3920	.49	*.48	.50	*.52	1.57	*.1.48
Age diff.	> 7	3089 7991	.49	*.48	.51	*.51	*1.58	1.48
	≤ 7	2076 3939	.49	*.51	.50	*.54	*1.56	1.48
Validators' Event Fam.	> 3	1386 540	.49	.44	.51	.47	*1.60	*1.42
	≤ 3	2099 676	.48	.45	.49	.48	*1.58	*1.47
Validators' Openness	+	2685 1472	.49	.49	.50	.52	1.57	1.47
	–	3000 1568	.49	.48	.50	.51	1.57	1.48
Validators' Conscien.	+	3151 1638	*.48	.51	*.49	.53	*1.57	*1.49
	–	2589 1426	*.50	.51	*.51	.54	*1.56	*1.46
Validators' Extraversion	+	2878 1685	.49	*.48	.50	*.51	*1.58	*1.51
	–	2812 1535	.50	*.52	.51	*.55	*1.56	*1.46
Validators' Agreeabl.	+	2675 1451	.49	*.51	.51	*.54	*1.58	1.47
	–	2930 1553	.48	*.45	.49	*.49	*1.56	1.47
Validators' Emot. Stab.	+	2838 3009	*.48	*.48	*.49	*.51	*1.57	*1.50
	–	2792 2897	*.50	*.51	*.51	*.54	*1.56	*1.46

- Validators agree more with each other than with the generator
- V–G agreements:
  - Higher agreement for Female pairs
  - Low age difference leads to higher agreement
- V properties only:
  - Event familiarity hurts agreement for appraisal
  - We expected Open annotators to perform better.
  - Emotional stability “hurts” emotion annotation.
  - Extraversion, Conscient., Agreeableness help.
- Most differences are quite small (but significant)

Examples (writer/reader/avg. writer–reader agreement as error)

- All writers/readers agree on emotion, high average appraisal agreement

  pride, .65 — I baked a delicious strawberry cobbler
  fear, .84 — A housemate came at me with a knife
- All writers/readers agree on emotion, low average appraisal agreement

  disgust, 2.0 — His toenails where massive
  fear, 2.1 — I felt ... going in to hospital
- All readers agree on the emotion, but not with the writer, high appraisal agreement

  trust, joy, .87 — I am with my friends
  anger, fear, 1.1 — My waters broke early during pregnancy
- All readers agree on the emotion, but not with the writer, low appraisal agreement

  pride, sadness, 1.7 — That I put together a funeral service for my Aunt

Appraisals add additional information to emotion analysis

That I put together a funeral service for my Aunt

Dimension	Writer	Readers	△
Emotion	Pride	Sadness	
Suddenness	4	3.6	0.4
Familiarity	1	2.0	−1.0
Predictability	1	1.8	−0.8
Pleasantness	4	1.0	3.0
Unpleasantness	2	4.8	−2.8
Goal-Relevance	4	2.6	1.4
Chance-Resp.	4	4.4	−0.4
Self-Resp.	1	1.2	−0.2
Other-Resp.	1	1.4	−0.4
Conseq.-Predict.	2	1.8	0.2
Goal Support	1	1.2	−0.2
Urgency	2	3.8	−1.8
Self-Control	5	3.2	1.8
Other-Control	3	2.0	1.0
Chance-Control	1	4.6	−3.6
Accept-Conseq.	4	2.4	1.6
Standards	1	2.4	−1.4
Social Norms	1	1.2	−0.2
Attention	4	4.4	−0.4
Not-Consider	1	3.8	−2.8
Effort	4	4.6	−0.6

Emotion Annotation Result

Conclusion

Annotators can quite well reconstruct authors emotion,
but there is a small and significant agreement drop.

Challenge

Authors recall "important" events. We do (presumably) not get a realistic subsample of event descriptions as they appear in the wild.

- Not shown: appraisals help to disambiguate emotion categories in automatic models

Potential Reason for V–G Discrepancy

- Isolated events are not sufficient
- Subjectivity is not only personality and demographics

Idea: Generate backstories to explain emotions/appraisals

Event

"The loudspeaker suddenly malfunctioned and went silent."

- Many emotion interpretations possible.

⇒ We autogenerate stories that explain such event for a given emotion.

J. Schäfer et al. (2025). Shaping Event Backstories to Estimate Potential Emotion Contexts. arXiv: 2508.09954 [cs.CL]. URL: https://arxiv.org/abs/2508.09954

Backstories

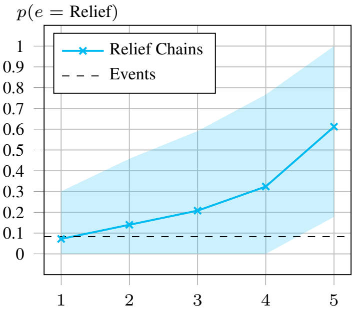
Relief: I was tasked with giving a presentation to a large crowd. The sound system malfunctioned, amplifying my voice to an ear-piercing level. The sound technician ignored the problem and chatted with someone. The audience covered their ears and looked at me with discomfort. The loudspeaker suddenly malfunctioned and went silent.

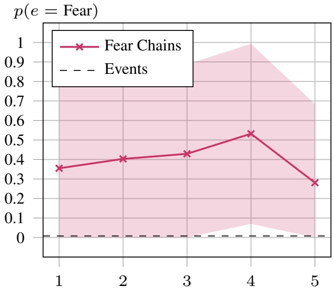
Fear: I arrived at a remote wilderness survival training camp, where the instructors emphasized the importance of following loudspeaker instructions for safety. The instructors warned us about the toxic waste site nearby and explained that the loudspeaker would alert us to any changes in air quality. During the first exercise, I struggled to navigate the challenging terrain, but the loudspeaker provided crucial guidance, helping me stay on track. I completed a difficult obstacle course, relying heavily on the loudspeaker's instructions to avoid hazards and find the safest route. The loudspeaker suddenly malfunctioned and went silent.

Pride: I spent the entire morning upgrading the sound system with a new backup system to prevent technical issues. The event organizer informed me that the conference was running 30 minutes behind schedule, giving me extra time to test the new backup system. I used the extra time to run a series of tests on the sound system, trying to simulate potential failures. The keynote speaker began to talk, and the sound system was working flawlessly, but I was still waiting for a real test of the new backup system. The loudspeaker suddenly malfunctioned and went silent.

Results in a Nutshell

- Backstories make interpretation more clear for models and annotators (details not shown for time reasons).
- Effect more pronounced for some emotions than others





Summary and Next Steps

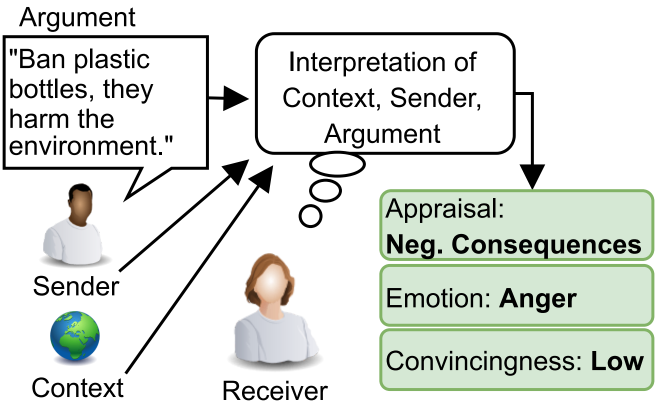
- We learned about appraisals as an emotion model that links the evaluation of events and emotions.
- It explains emotion categories, but also acts as a model in itself.
- Sometimes, it might just be the more appropriate emotion model.

Outline

Argument Convincingness

- Argument quality includes:
  - Logical structure: Logos
  - Speaker credibility: Ethos
  - Emotional appeal: Pathos
- Arguments are subjectively evaluated

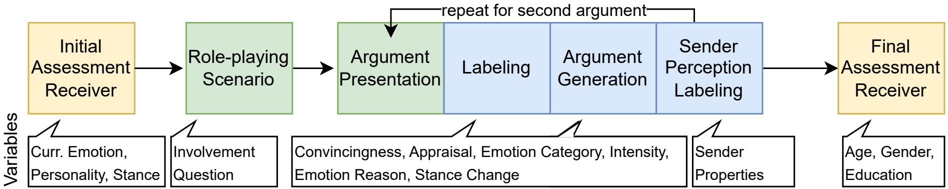
The Contextualized Argument Appraisal Framework



Do appraisals explain convincingness?

L. Greschner et al. (2025). Trust Me, I Can Convince You: The Contextualized Argument Appraisal Framework. to be on arxiv soon. arXiv: 0000.00000 [cs.CL]. URL: https://arxiv.org/abs/0000.00000

Argument Appraisal Annotation Framework



- 39 topics
- 800 arguments
- Each 5 annotations
- 9,404 £

Argument Appraisal Variables

Dimension	Description
Suddenness	the argument appears sudden or abrupt to the receiver
Suppression	the receiver tries to shut the argument out of their mind
Familiarity	the argument is familiar to the receiver
Pleasantness	the argument is pleasant for the receiver
Unpleasantness	the argument is unpleasant for the receiver
Consequencial Importance	the argument has important consequences for the receiver
Positive Consequentiality	the argument has positive consequences for the receiver
Negative Consequentiality	the argument has negative consequences for the receiver
Consequence Manageability	the receiver can easily live with the unavoidable consequences of the argument
Internal Check	the consequences of the argument clash with the receiver's standards and ideals
External Check	the consequences of the argument violate laws or socially accepted norms
Response urgency	the receiver urges to immediately respond to the argument
Cognitive Effort	processing the argument requires a great deal of energy of the receiver
Argument Internal Check	statements in the argument clash with the receiver's standards and ideals
Argument External Check	statements in the argument violate laws or socially accepted norms

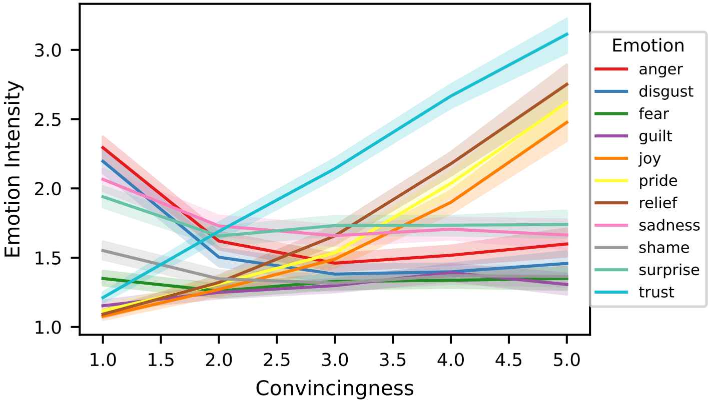
Convincingness and Emotions – Average Values

	Suddenness	Suppression	Familiarity	Pleasantness	Unpleasantness	Consequencial Importance	Positive Consequentiality	Negative Consequentiality	Consequence Manageability	Internal Check	External Check	Response urgency	Cognitive Effort	Argument Internal Check	Argument External Check
Anger	2.4	1.9	2.6	1.2	3.0	1.9	1.6	1.6	2.4	2.8	2.7	2.4	3.7	2.0	3.1
Disgust	2.4	2.1	2.5	1.2	3.2	1.8	1.5	1.7	2.3	2.8	2.9	2.6	3.8	1.8	3.2
Fear	1.9	1.6	2.8	1.6	2.5	2.2	2.1	1.8	2.4	2.3	2.0	1.9	3.5	1.9	2.0
Guilt	1.7	1.5	2.8	1.6	2.3	1.9	1.7	1.6	2.3	2.2	1.9	1.5	3.6	1.6	1.9
Joy	1.4	1.1	3.1	3.5	1.2	2.0	2.5	2.2	2.1	2.5	1.3	1.5	3.1	1.6	1.4
Pride	1.4	1.2	3.3	3.2	1.3	2.0	2.5	2.1	2.2	2.4	1.4	1.6	3.4	1.4	1.4
Relief	1.4	1.1	3.2	3.0	1.3	2.0	2.5	2.0	2.5	2.3	1.4	1.5	3.5	1.4	1.4
Sadness	1.9	1.5	2.9	1.3	2.7	1.8	1.6	1.6	2.3	2.7	2.2	2.0	3.6	1.7	2.5
Shame	2.0	1.6	2.8	1.4	2.8	2.0	1.7	1.9	2.5	2.5	2.2	2.1	3.5	1.9	2.4
Surprise	2.1	1.4	2.1	1.7	1.9	1.8	1.6	1.5	2.3	2.5	1.9	1.8	3.5	1.7	2.1
Trust	1.4	1.1	3.0	2.5	1.3	1.9	2.0	1.6	2.4	2.2	1.3	1.4	3.6	1.3	1.4

- Generally high cognitive effort
- External Check explains anger and disgust
- Familiarity indicative for positive emotions

Correlations of Emotions with Convincingness

Emotion	r
Trust	0.570
Relief	0.511
Pride	0.458
Joy	0.435
Guilt	0.105
Fear	0.006
Surprise	−0.072
Shame	−0.073
Sadness	−0.153
Anger	−0.265
Disgust	−0.264



Correlations of Appraisals with Convincingness

Appraisal	r
Pleasantness	0.566
Positive Consequentiality	0.392
Familiarity	0.327
Negative Consequentiality	0.203
Consequential Importance	0.141
Consequence Manageability	−0.034
Cognitive Effort	−0.061

Appraisal	r
Internal Check	−0.103
Argument Internal Check	−0.109
Response Urgency	−0.242
Suppression	−0.326
Suddenness	−0.342
External Check	−0.355
Unpleasantness	−0.385
Argument External Check	−0.497

- Pleasant arguments whose outcomes are good for the self and which are familiar are more convincing.
- Surprising arguments and those which go against laws or social standards are less convincing (and cause anger and disgust).

Outline

Introduction

- Synthetic data creation has advantages:
  - Direct access to the author's assessment
  - Privacy: authors are aware what they share and can filter
- Potential issues:
  - Data is not realistic
  - People recall particularly "prototypical" events
  - Type of data might differ due to missing post creation triggers

E. Troiano et al. (2023). "Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction". In: Computational Linguistics 49.1

Approach: Data elicitation strategies

- Creation:
  - “Think of an event that caused an emotion X in you.”
  - “Write a social media post text about that.”
  - “Select an image you want to share from a CC image data base.”
- Donation:
  - “Pick a multimodal post from your social media timeline that you made because the associated event caused emotion X.”
  - “Copy paste the text and the image.”
- Recent:
  - “Pick the 10 most recent posts from your social media timeline.”
  - “Annotate them for the following emotion set.”

Data Example

Name
@Username

Absolutely insane, what is going on?!

Creation post labeled as surprise.

Name2
@Username2

Exhibit 2

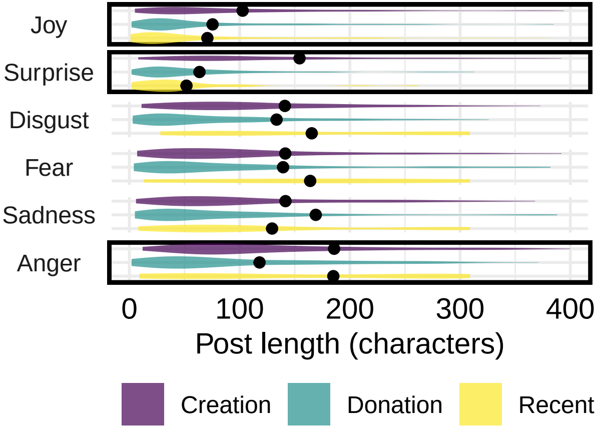
Trump supporters say ear bandages are 'sign of love'

Several supporters of former President Donald Trump wore bandages on their ears to the third night of the Republican National Convention (RNC) in Milwaukee, Wisconsin.

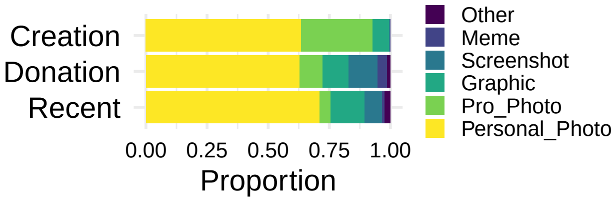
Members of the RNC's Arizona delegation said they were wearing the bandages as a sign of solidarity with the former president after he survived an assassination attempt.

Recent post labeled as anger.

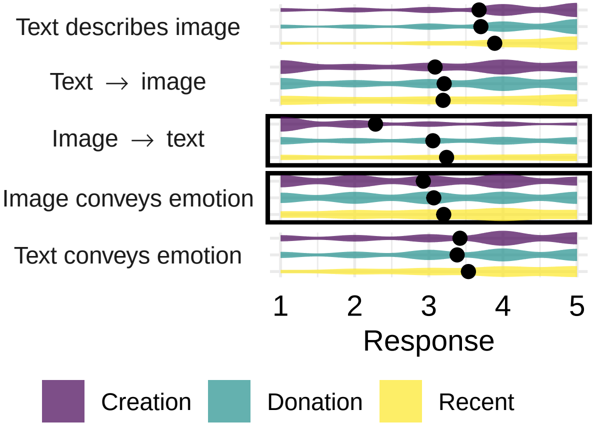
Are the subcorpora comparable? – Post Length



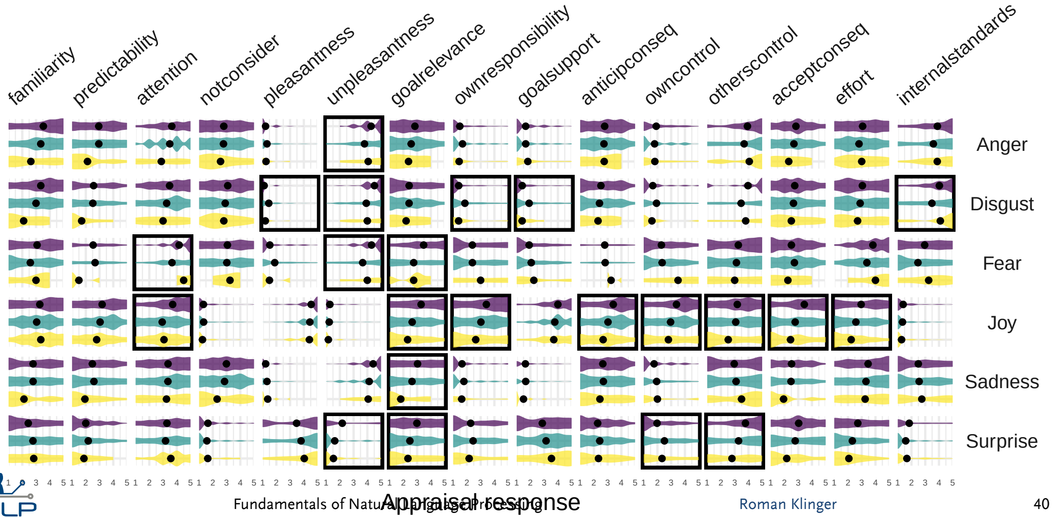
Are the subcorpora comparable? – Image Type



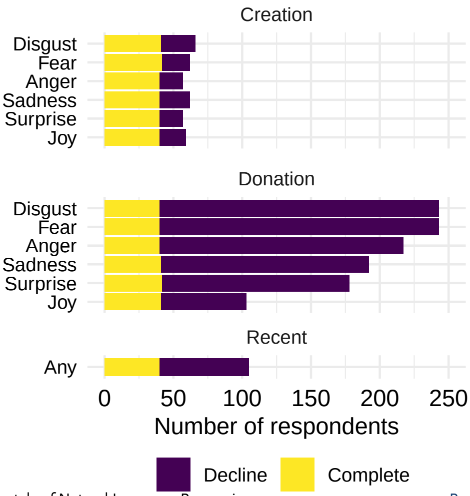
Are the subcorpora comparable? – Text–Image Relation



Are the subcorpora comparable? – Appraisal–Emotion



Are the subcorpora comparable? – Participant acceptance



Are the differences a problem?

Experiment

- Fine-tune RoBERTa with CLIP/early fusion to predict emotions
  - Train on Donation vs. train on Creation

Results

- No big performance differences: F score .38 vs. .40
  ⇒ The experimentally elicited data is fine to optimize a model.
- But: The estimate on donated data is overall optimistic!
  F score of .60 and .62.
  ⇒ Real data is required to estimate model performance.
- Zero-Shot prompting (Llama3.2-vision) leads to slightly better results for donated data.

Outline

Summary

- Appraisals are an emotion model that explain the cognitive evaluation process that is part of an emotion
- Appraisals can be annotated and modeled
- ...but they are subjective and require context
- Appraisals are a informative approach to explain argument convincingness
- Experimentally elicited data is fine for model training, but we need real data for performance estimation
  (shown for emotion categories only so far, though)

Thank you for
your attention.
Questions? Remarks?

?

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Thanks to

- Enrica Troiano
- Laura Ana Maria Oberländer née Bostan
- Lynn Greschner
- Johannes Schäfer
- Sabine Weber
- Christopher Bagdon
- Carina Silberer
- Kai Sassenberg
- All of BamNLP

University of Bamberg

Appraisals as Emotion Model in NLP

Why we need them and how to acquire data

Workshop, Ghent, September 26, 2025

Roman Klinger
roman.klinger@uni-bamberg.de

romanklinger.de romanklinger
https://www.bamberg.de/nlproc/