

Psychological Concepts Challenge Natural Language Processing

Belief and Facts, Emotions and Appraisal

Linguistische Werkstatt, May 15, 2024

Roman Klinger
roman.klinger@uni-bamberg.de

 @roman_klinger  romanklinger
<https://www.bamberg.de/nlproc/>

Example Text

“Thank you for inviting me to give this talk. It’s my pleasure to be here, despite being a bit nervous to talk in front of linguistics as a non linguist. It’s actually the first time that I do that, ever, I think.”

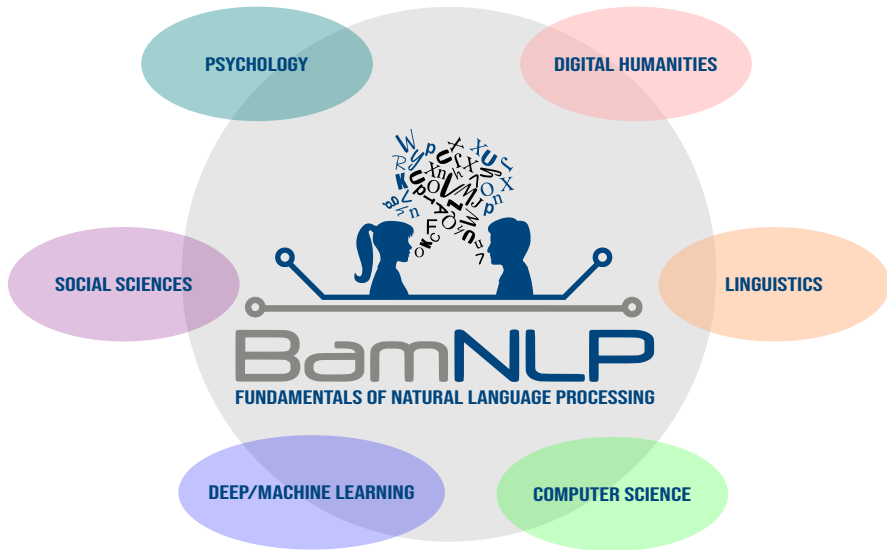
- What do you learn about my current emotional state from this text?
- Do you believe that this is actually true?



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



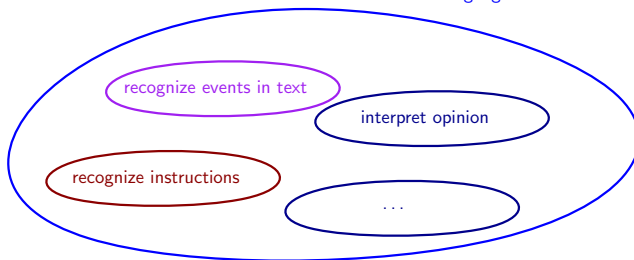




Natural Language Processing Tasks

What does natural language processing research look like?

natural language textual communication



- NLP research does barely attempt to solve everything that humans can do.
- Instead: predefined (narrow) tasks.
- Some tasks are established and well defined.
- Others are still in the process of formalization.

• We will now look at a couple of examples.



Example Task: Named Entity Recognition

Example Input (one of many) to Instruct an Automatic Machine Learning Model

Input: Both Gabriele Knappe and Stefanie Stricker work at the Uni Bamberg.

Output: Gabriele Knappe ; Stefanie Stricker

Application

Input: Roman Klinger works at the University of Bamberg.

Output: Roman Klinger

- I specified the task with an example (standard machine learning setup: supervised learning).
- An alternative task specification would be an instruction: “Annotate all person names.”



Example Task: Machine Translation de-en

Example

Input: Roman Klinger arbeitet an der Uni Bamberg.

Output: Roman Klinger works at the University of Bamberg.



Example Task: Conditional Text Generation

Example

Input: “When he walked into the restaurant”, Joy

Output: “he was delighted to see that his husband was already there.”



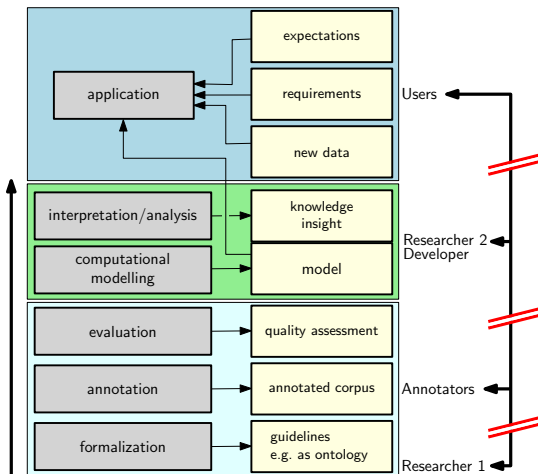
Example Task: Natural Language Inference

Example

- Input: “A soccer game with multiple males playing.”;
“Some men are playing a sport.”
- Output: entailment
- Input: “A man inspects the uniform of the person.”;
“The man is sleeping.”
- Output: contradiction



Natural Language Processing Research



- How to formalize a concept without inappropriately simplifying it, while making it “computable”?
- How to setup the annotation task such that it leads to reliable text assessments?
- How to model concept properties correctly such that annotations can be automatized?
- Do models generalize?
Are users happy?



Annotation Challenges

Questions

- Is the task objectively decidable?
(entities vs. entailment or translation)
- Is the text alone sufficient to solve the task or is more context needed?
(textual entailment vs. multimodal data or author profiling)
- Is it a classification or regression task?
(emotion classification vs. arousal regression)

Implications

- Do we have access to the context? How much to show?
- Show isolated instance or request comparative annotations?
- Carefully train annotation experts or do crowdsourcing?

Modeling



Find a function that takes

- **text** (and **additional information**) as input
- and automatically predicts **output/annotation**.



Modeling Approaches

- Rule-based methods, lexicon-based approaches
 - + Transparent
 - + Can be well grounded in theories
 - Often conceptually too simple
 - Difficult to achieve good performance
- Machine Learning/Deep Learning, Supervised or via Reinforcement Learning
 - + Learns the task from data
 - + No need to fully specify the task manually
 - SOTA: Fine-tuning a pretrained language model
 - Data is required
 - Prone to overfitting to data
- Prompting, Prompt Learning; Learning from Instructions
 - + Potentially good generalization, potentially only needs few example instances
 - Needs a large (instruction-tuned) language model



Prompting with Instruction-tuned

Step 1: Train a model to understand language: Language modeling objective

- Input: “I want to eat” — Output: “Spaghetti”.
- Observation: Input/Output pairs can be created without human supervision!

Step 2: Fine-tune this model to solve instructions

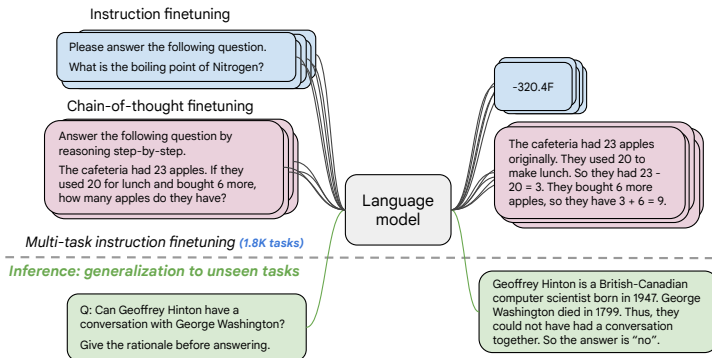
- Input: “Classify the sentiment: ‘I like the company’” — Output: “Positive”.
- Obs.: We need many tasks & huge models to achieve generalization across tasks.

Step 3: Fine-tune with reinforcement learning from human-feedback on unseen tasks

- Given a human input and a model’s output, let a human judge it’s quality.
- Observation: We need many humans to do that.

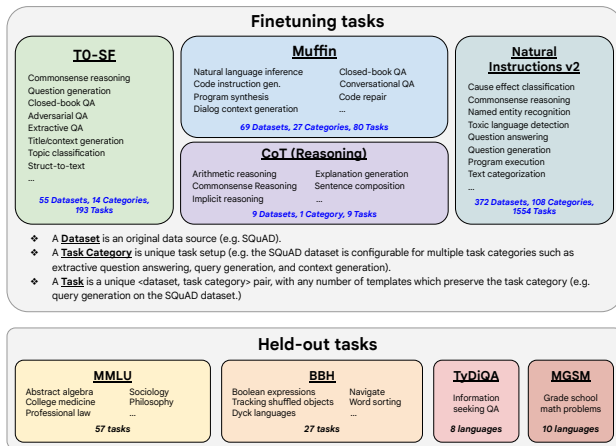


Example: Flan-T5 (1)





Example: Flan-T5 (2)



Outline

- 1 NLP Research Methods
- 2 Emotion Analysis
- 3 Appraisal-based Emotion Analysis
- 4 Deception Detection
- 5 Take Home

Outline

1 NLP Research Methods

2 Emotion Analysis

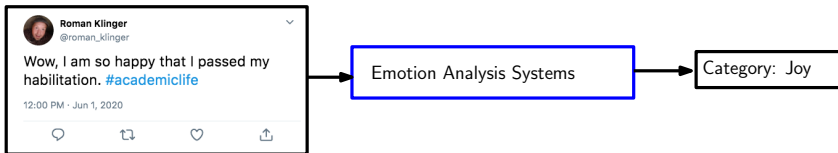
3 Appraisal-based Emotion Analysis

4 Deception Detection

5 Take Home

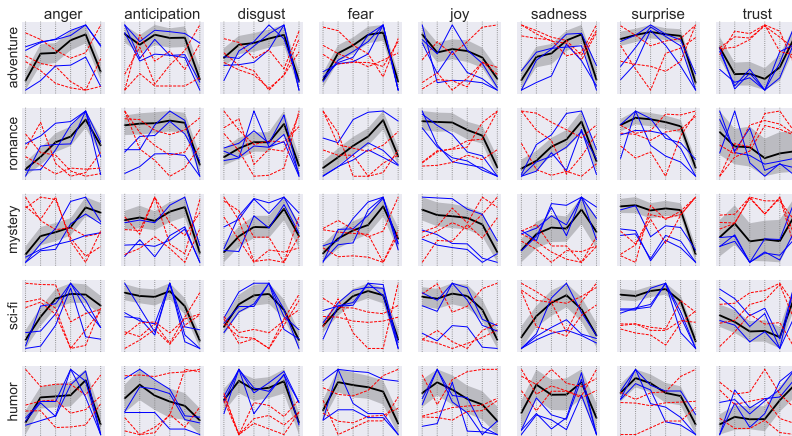


Emotion Analysis: What we want to do.





Literary Studies



Kim et al., 2017.

Investigating the Relationship between Literary Genres and Emotional Plot Development. LaTeCH@ACL



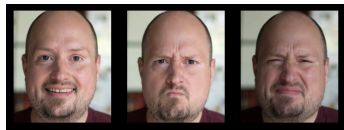
Dominant Emotions Expressed in News Articles

Emotion	Dominant Emotion
Anger	The Blaze, The Daily Wire, BuzzFeed
Annoyance	Vice, NewsBusters, AlterNet
Disgust	BuzzFeed, The Hill, NewsBusters
Fear	The Daily Mail, Los Angeles Times, BBC
Guilt	Fox News, The Daily Mail, Vice
Joy	Time, Positive.News, BBC
Love	Positive.News, The New Yorker, BBC
Pessimism	MotherJones, Intercept, Financial Times
Neg. Surprise	The Daily Mail, MarketWatch, Vice
Optimism	Bussines Insider, The Week, The Fiscal Times
Pos. Surprise	Positive.News, BBC, MarketWatch
Pride	Positive.News, The Guardian, The New Yorker
Sadness	The Daily Mail, CNN, Daily Caller
Shame	The Daily Mail, The Guardian, The Daily Wire
Trust	The Daily Signal, Fox News, Mother Jones

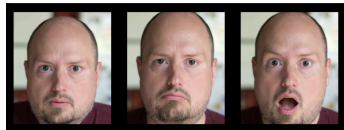
Bostan et al., 2020.

GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. LREC

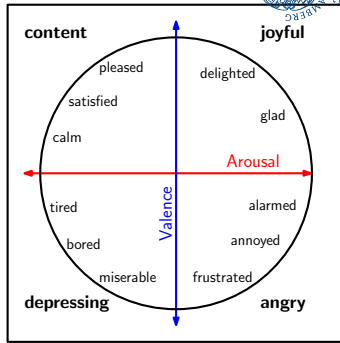
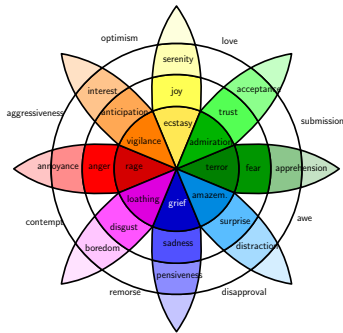
How to define a categorical system of emotions?



Joy Anger Disgust



Fear Sadness Surprise



- Emotion models in psychology explain how emotions are developed.
- Text analysis models learn to associate textual realizations to emotion concepts. They do not (explicitly?) use knowledge from such theories.

Outline

- 1 NLP Research Methods
- 2 Emotion Analysis
- 3 Appraisal-based Emotion Analysis
- 4 Deception Detection
- 5 Take Home



Emotion Examples

Which emotion is associated with the examples?

How did you recognize that?

- “She became angry.”
- “A tear is running down his face.”
- “We are going for a walk at the beach.”
- “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.

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** ...”

		
Feeling	Expression	Bodily Symptom
Action Tendency	Cognitive Appraisal	
Fear		

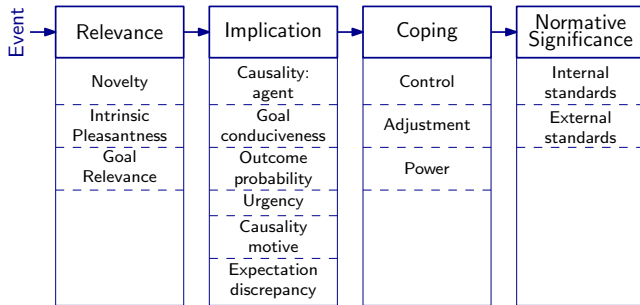
Event

Components

Name



Cognitive Appraisal in Scherer's Component Process model



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



Research Questions

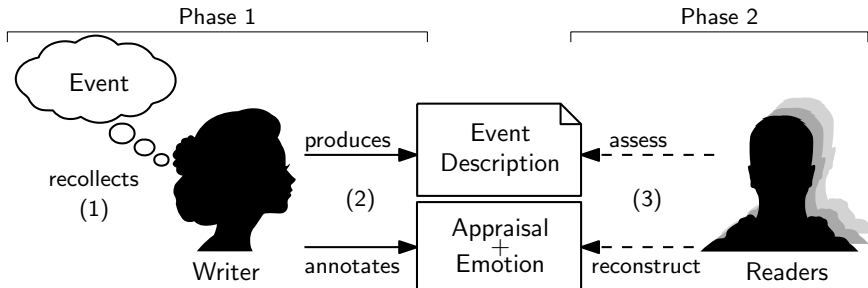
- Can appraisals be annotated reliably?
- Can we predict appraisal variables from event descriptions?
- Do appraisals help emotion categorization?

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



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



Questions and Answers

- Do readers agree more with each other than with the writers?
(does the writer make use of information that the readers do not have)
 - Yes, a bit for emotions; clearly for the appraisals.
- Does it matter if annotators share demographic properties?
 - Females agree more with each other, but men less.
 - People of similar age agree more.
- Does personality matter?
 - Extraverted, conscientious, agreeable annotators perform better.

Setup:

- Filter instances for attribute, compare with F_1 /RMSE
- Significance test with bootstrap resampling for .95 confidence interval



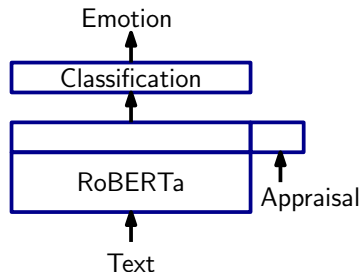
Examples (writer/reader/avg. writer–reader agreement as error)

- All writers/readers agree on emotion, **high** average appraisal agreement
pride, .65
fear, .84
I baked a delicious strawberry cobbler
A housemate came at me with a knife
- All writers/readers agree on emotion, **low** average appraisal agreement
disgust, 2.0
fear, 2.1
His toenails where massive
I felt ... going in to hospital
- All readers agree on the emotion, but **not with the writer**, **high** appraisal agreement
trust, joy, .87
anger, fear, 1.1
I am with my friends
My waters broke early during pregnancy
- All readers agree on the emotion, **but not with the writer**, **low** appraisal agreement
pride, sadness, 1.7
shame, relief, 1.8
That I put together a funeral service for my Aunt
I tasked with sorting out some files from the office the
previous day and I slept off when I got home



Modeling Results

- Classification with RoBERTa-based models
 - Appraisal Classification: 75 F_1
 - Emotion classification: 59 F_1
 - + Appraisals: +2pp F_1
(+10 for guilt, +6 for sadness)
- ⇒ Appraisals help to build better models.





Examples where Appraisals correct the Emotion Classifier

- When my child settled well into school
- broke an expensive item in a shop accidentally
- my mother made me feel like a child
- I passed my Irish language test
- His toenails where massive

trust→relief

guilt→shame

shame→anger

pride→relief

pride→disgust



Conclusion & Summary

- We presented the first self-annotated large-scale appraisal corpus
- Annotators can reliably recover both emotions and appraisals (demographics play a significant but small role)
- Appraisals help emotion categorization for some emotion categories
- More importantly: Appraisals help to understand reasons for disagreement

Outline

- 1 NLP Research Methods
- 2 Emotion Analysis
- 3 Appraisal-based Emotion Analysis
- 4 Deception Detection
- 5 Take Home



Deception

Deception

The term “**deception**” refers to the intentional act of causing someone to hold a false belief, which the deceiver knows to be false or believes to be untrue.

Examples: Lies, exaggerations, omissions

A. Velutharambath, A. Wührl, et al. (2024). “Can Factual Statements be Deceptive? The DeFaBel Corpus of Belief-based Deception”. In: LREC-COLING

Linguistic Cues of Deception



- Deceptive statements have fewer self-references
- More ambiguous statements
- Longer sentences, more details
- Readability is lower



Cross-Corpus Deception Detection

- Do linguistic properties hold across corpora?
- Do models generalize from one corpus to another?

A. Velutharambath and R. Klinger (2023). "UNIDECOR: A Unified Deception Corpus for Cross-Corpus Deception Detection". In: WASSA



Cross-Corpus Deception Detection

Dataset	Domain	Truthful	Deceptive	Total	TC	SC
Bluff the listener (BLUFF)	game	251 (33.3%)	502 (66.7%)	753	241.66	11.5
Diplomacy dataset (DIPLOMACY)	game	16402 (94.9%)	887 (5.1%)	17289	24.53	1.7
Mafiascum dataset (MAFIASCUM)	game	7439 (76.9%)	2237 (23.1%)	9676	4690.69	362.8
Multimodal Decep. in Dialogues (BOXOFLIES)	game	101 (20.2%)	400 (79.8%)	501	12.2	1.6
Miami University Decep. Detection Db. (MU3D)	interview	160 (50.0%)	160 (50.0%)	320	131.7	5.7
Real-life trial data (TRIAL)	interview	60 (49.6%)	61 (50.4%)	121	79.85	3.9
Cross-cultural deception (CROSSCULTDE)	opinion	600 (50.0%)	600 (50.0%)	1200	80.0	4.5
Deceptive Opinion (DECOP)	opinion	1250 (50.0%)	1250 (50.0%)	2500	65.56	4.0
Boulder Lies and Truth Corpus (BLTC)	review	1041 (69.8%)	451 (30.2%)	1492	116.92	6.5
Deception in reviews (DEREV2014)	review	118 (50.0%)	118 (50.0%)	236	145.22	6.7
Deception in reviews (DEREV2018)	review	1552 (50.0%)	1552 (50.0%)	3104	176.6	8.1
Deceptive opinion spam (OPSPAM)	review	800 (50.0%)	800 (50.0%)	1600	170.5	9.5
Online deceptive reviews (ONLINEDE)	review	101431 (85.9%)	16694 (14.1%)	118125	171.5	7.2
Open Domain Deception (OPENDOMAIN)	statement	3584 (50.0%)	3584 (50.0%)	7168	9.33	1.0
		134789 (82.1%)	29296 (17.9%)	164085	436.88	31.05



Cross-Corpus Deception Detection

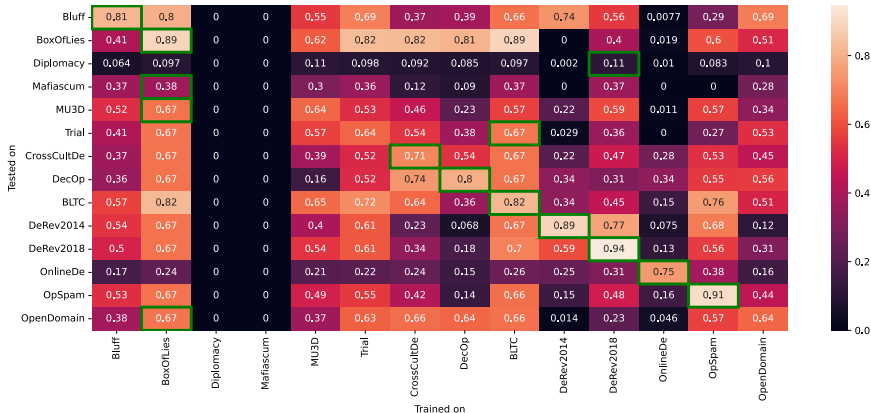
Features	Datasets													
	BLTC	BLUFF	BOXOFLIES	CROSSCULTDE	DECOP	DEREV2014	DEREV2018	DIPLOMACY	MAFIASCUM	MU3D	ONLINEDE	OPENDOMAIN	OPSPAM	TRIAL
Analytic	.13	-.04	.12	.01	.02	-.25	.23	.02	-.02	.14	.10	.05	.15	.25
Authentic	.03	-.05	.00	.28	.22	.28	-.05	-.03	-.02	.07	.00	-.04	-.09	-.09
BigWords	.02	.00	.18	.04	.05	-.21	.24	.01	-.01	.18	-.01	.03	-.08	.09
Clout	.00	.00	.02	-.11	-.28	-.45	.00	.02	.02	.03	-.05	.01	.10	.26
Cognition	-.08	.17	-.05	.02	.07	-.06	-.13	-.01	-.01	-.17	.00	-.09	-.06	-.28
GunningFog	.18	-.21	.12	.21	.25	.01	.13	-.09	-.03	-.04	.13	.02	.02	.06
Kincaid	.18	-.21	.14	.2	.24	.01	.13	-.08	-.03	-.04	.13	.03	.02	.06
Linguistic	-.07	.10	-.15	.04	.10	.29	-.14	-.02	-.03	-.16	-.05	-.05	-.18	-.08
Period	.01	-.07	.02	-.11	-.18	.26	-.07	.00	.00	.03	.01	.03	.24	-.06
Physical	.02	.03	.15	-.04	-.16	-.25	.06	.00	.03	.04	-.15	-.01	-.01	.06
WC	.18	-.21	.04	.22	.25	.02	.13	-.10	.01	-.04	.13	-.02	.02	.06
auxverb	-.08	.12	-.06	-.08	-.09	.22	-.12	-.01	.02	-.15	.00	.03	-.08	-.21
focusfuture	-.09	.09	-.02	-.04	-.08	-.17	-.2	-.01	.02	-.04	.01	-.04	-.16	.08
function	-.05	.13	-.03	.00	.10	.25	-.06	-.04	-.03	-.15	-.03	-.05	-.23	-.23
i	-.06	-.15	-.07	.13	-.3	.39	-.16	-.05	.02	-.01	-.12	-.04	-.33	-.13
shehe	.01	-.11	-.03	-.15	.00	-.17	-.07	.00	-.04	-.14	.04	-.04	-.01	-.18
verb	-.11	.07	-.09	-.06	-.07	.16	-.26	-.02	.00	-.14	-.07	-.01	-.16	-.14
you	-.10	.17	-.03	-.05	-.07	-.19	-.23	.01	.03	-.08	-.05	-.05	.01	-.05

- We cannot find a consistent property of deception across corpora.



Cross-Corpus Deception Detection

Within-corpus and cross-corpus results for RoBERTa



Model does not generalize across corpora.

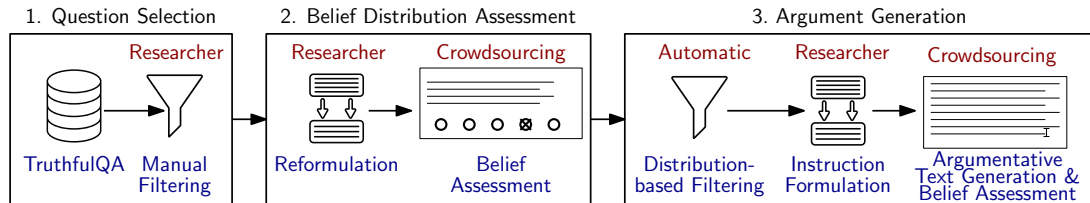
Research Hypotheses



- Something is wrong here...
- We assume that model's mostly learn topic/domain specific properties of lies.



Belief-based Deception Framework and Corpus (DeFaBel)



“Wenn man einen Regenwurm durchschneidet, entstehen zwei Regenwürmer” – Who believed it?



Ein Regenwurm hat im Gegensatz zu ändern Tieren oder Säugetieren kein Gehirn sondern ein dezentrales Nervensystem, welches seine Funktionen steuert. Ebenso hat er kein Herz oder andere singuläre Organe, die für ihr lebenswichtig sind. Verdauung, Atmung sind nicht an einen Ort gebunden. Das führt dazu, dass ein durchgeschnittener Regenwurm zwei Teile bildet, die unabhängig voneinander lebensfähig sind. Nach einer gewissen Zeit, wachsen an den Enden jeweils Schwanz/Kopf, die mit den ursprünglichen Enden des Wurm vergleichbar sind - es sind zwei neue, lebensfähige Regenwürmer entstanden.

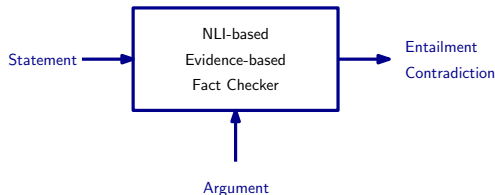
Schneidet man einen Regenwurm durch, so verdoppelt sich das Tier sozusagen, weil sich die beiden Hälften des durchgeschnittenen Wurmes zu eigenständigen Wesen entwickelt. Das liegt daran, dass der Regenwurm ein verblüffend komplexes Wesen ist. Er hat die Fähigkeit, seine inneren Organe, sein Herzkreislaufsystem und sein Gehirn bei Bedarf zu duplizieren. Das liegt in der Entwicklungsgeschichte des Regenwurms begründet. So nützlich er im Garten ist, so leicht wird er auch vom Menschen aus Versehen geteilt. Das weiß jeder Gärtner, der im Übereifer beim Jäten schon einmal einen Regenwurm geteilt hat. Der Regenwurm hat sich in seiner Evolution diesen tragischen Unfällen angepasst, indem er die Fähigkeit entwickelt hat, sich bei Bedarf aus zwei Hälften neu entstehen zu lassen. Praktisch, oder?



Deception modeling in DeFaBel

Work in Progress:

- Current models do not recognize deception in this corpus
- We do not find the linguistic markers known to indicate deception in English
- But:
Deceptive arguments are less suitable to
fact-check the original statement than real arguments!





Take Home

- NLP Research is driven by task definitions, annotation and modeling
- Modeling emotions benefits from knowledge from psychological theories
- Deception (in German without topic bias) is not recognizable (yet)
- Deceptive arguments are less supportive for claims than honest ones

Thank you for
your attention.

Questions? Remarks?



Thanks to

- Ph.D. Students
 - Amelie Wührl
 - Aswathy Velutharambath
 - Yarik Menchaca Resendiz
 - Laura Oberländer
 - Enrica Troiano
- Collaborators
 - Kai Sassenberg

Funded by
DFG Deutsche
Forschungsgemeinschaft
German Research Foundation

Psychological Concepts Challenge Natural Language Processing

Belief and Facts, Emotions and Appraisal

Linguistische Werkstatt, May 15, 2024

Roman Klinger
roman.klinger@uni-bamberg.de

 @roman_klinger  romanklinger
<https://www.bamberg.de/nlproc/>