Universität Stuttgart
Institut für
Maschinelle Sprachverarbeitung

Natural Language Processing Tasks and Methods

Challenges for Emotion Analysis and Generation

ZPID, Dec 13, 2023

Roman Klinger
roman.klinger@ims.uni-stuttgart.de

@roman_klinger romanklinger
https://www.romanklinger.de/

https://www.romanklinger.de/talks/zpid2023.pdf

About Myself

About Myself

- 1999–2006: Studies at University of Dortmund:
  Computer science with minor psychology

UNIVERSITÄT DORTMUND
FACHBEREICH INFORMATIK

Roman Klinger

Komposition von Musik mit Methoden der Computational Intelligence

– Diplomarbeit –

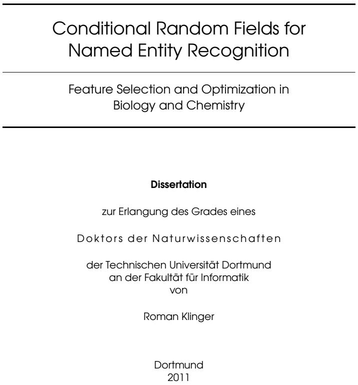
1. Juni 2006

Lehrstuhl 11
Computational Intelligence
Fachbereich Informatik
Universität Dortmund

Gutachter:
Prof. Dr. G. Rudolph
Dr. L. Hildebrand

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


Conditional Random Fields for
Named Entity Recognition

Feature Selection and Optimization in
Biology and Chemistry

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

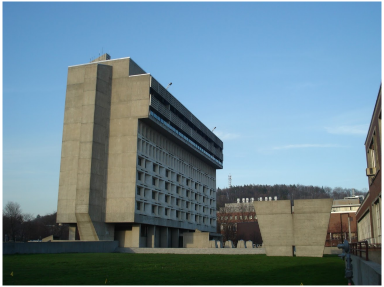
der Technischen Universität Dortmund
an der Fakultät für Informatik
von

Roman Klinger

Dortmund
2011

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



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

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

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

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

About Myself

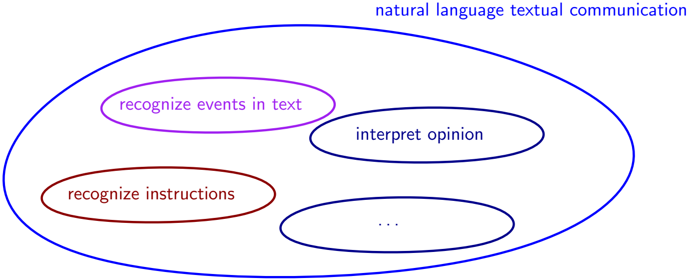
- 1999–2006: Studies at University of Dortmund:
  Computer science with minor psychology
- 2006–2010: Doctoral studies at Fraunhofer SCAI, St. Augustin:
  Biomedical text mining, machine learning
- 2010, 2013: Research visits at UMass Amherst:
  Probabilistic machine learning, MCMC inference
- 2011–2012: 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 Foundations of NLP, Uni Bamberg

Natural Language Processing Tasks

What does natural language processing research look like?

Natural Language Processing Tasks

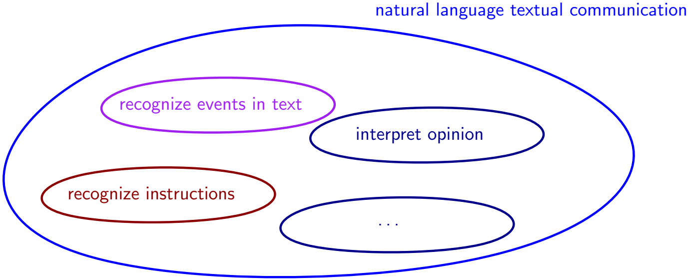
What does natural language processing research look like?



- NLP research does barely attempt to solve everything that humans can do.

Natural Language Processing Tasks

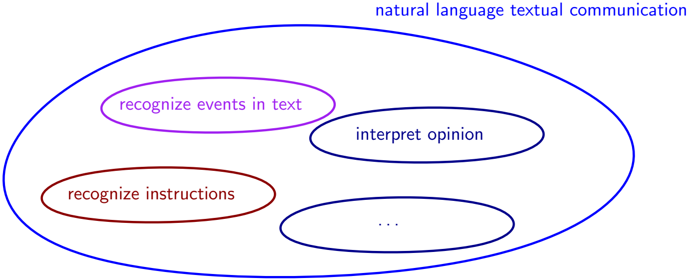
What does natural language processing research look like?



- NLP research does barely attempt to solve everything that humans can do.
- Instead: predefined (narrow) tasks.

Natural Language Processing Tasks

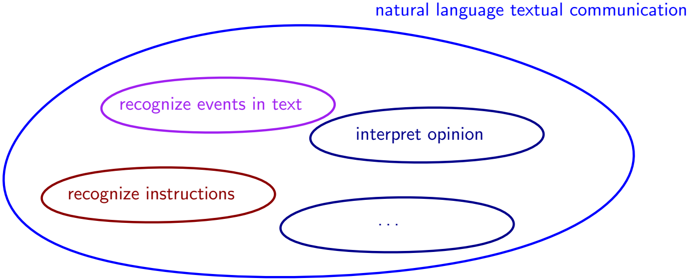
What does natural language processing research look like?



- NLP research does barely attempt to solve everything that humans can do.
- Instead: predefined (narrow) tasks.
- Some tasks are established and well defined.

Natural Language Processing Tasks

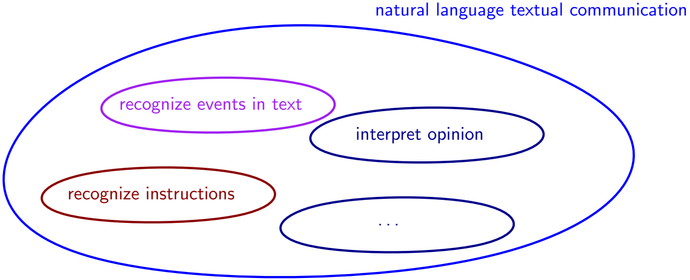
What does natural language processing research look like?



- 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.

Natural Language Processing Tasks

What does natural language processing research look like?



- 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: [redacted]

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

Input: Both Kai Sassenberg and André Bittermann work at the ZPID.
Output: Kai Sassenberg; André Bittermann

Example Task: [redacted]

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

Input: Both Kai Sassenberg and André Bittermann work at the ZPID.
Output: Kai Sassenberg; André Bittermann

Application

Input: Roman Klinger works at the University of Stuttgart.
Output:

Example Task: [redacted]

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

Input: Both Kai Sassenberg and André Bittermann work at the ZPID.
Output: Kai Sassenberg; André Bittermann

Application

Input: Roman Klinger works at the University of Stuttgart.
Output: Roman Klinger

Example Task: Named Entity Recognition

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

Input: Both Kai Sassenberg and André Bittermann work at the ZPID.
Output: Kai Sassenberg; André Bittermann

Application

Input: Roman Klinger works at the University of Stuttgart.
Output: Roman Klinger

Example Task: Named Entity Recognition

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

Input: Both Kai Sassenberg and André Bittermann work at the ZPID.
Output: Kai Sassenberg; André Bittermann

Application

Input: Roman Klinger works at the University of Stuttgart.
Output: Roman Klinger

- I specified the task with an example
  (standard machine learning setup: supervised learning).

Example Task: Named Entity Recognition

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

Input: Both Kai Sassenberg and André Bittermann work at the ZPID.
Output: Kai Sassenberg; André Bittermann

Application

Input: Roman Klinger works at the University of Stuttgart.
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: [redacted]

Example

Input: Roman Klinger arbeitet an der Uni Stuttgart.
Output: Roman Klinger works at the University of Stuttgart.

Example Task: Machine Translation de-en

Example

Input: Roman Klinger arbeitet an der Uni Stuttgart.
Output: Roman Klinger works at the University of Stuttgart.

Example Task: [redacted]

Example

Input: "When he walked into the restaurant", Joy
Output: "he was delighted to see that his husband was already there."

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: [redacted]

Example

- Input: “A soccer game with multiple males playing.”;
  “Some men are playing a sport.”
- Output: yes

- Input: “A man inspects the uniform of the person.”;
  “The man is sleeping.”
- Output: no

Example Task: Natural Language Inference

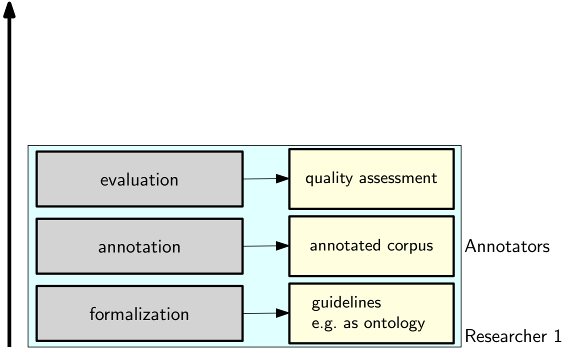
Example

- 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

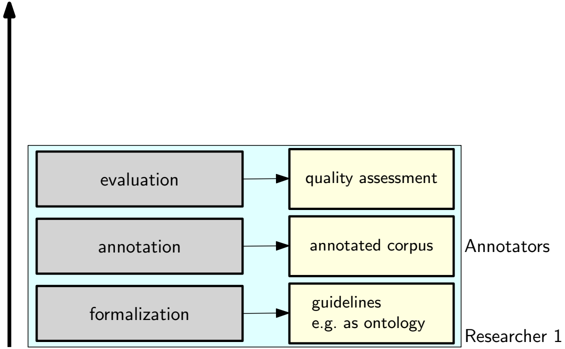
Natural Language Processing Research

Natural Language Processing Research



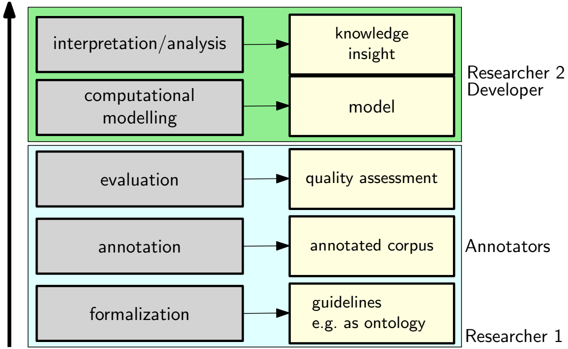
- How to formalize a concept without inappropriately simplifying it, while making it "computable"?

Natural Language Processing Research



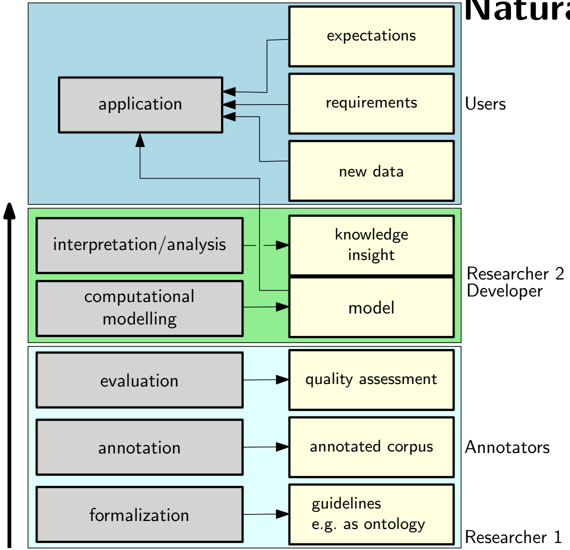
- How to formalize a concept without inappropriately simplifying it, while making it "computable"?
- How to setup the annotation task such that it leads to reliable text assessements?

Natural Language Processing Research



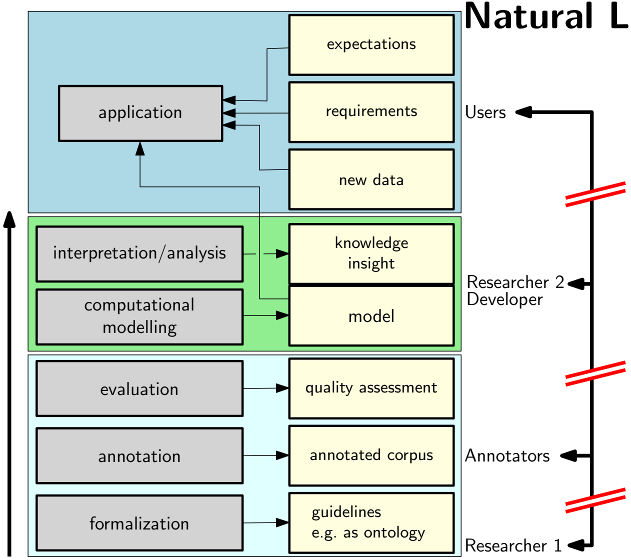
- How to formalize a concept without inappropriately simplifying it, while making it "computable"?
- How to setup the annotation task such that it leads to reliable text assessements?
- How to model concept properties correctly such that annotations can be automatized?

Natural Language Processing Research



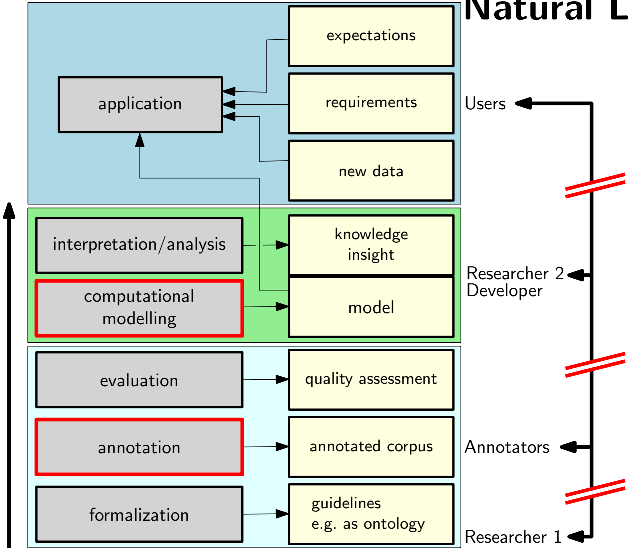
- How to formalize a concept without inappropriately simplifying it, while making it "computable"?
- How to setup the annotation task such that it leads to reliable text assessements?
- How to model concept properties correctly such that annotations can be automatized?
- Do models generalize?
  Are users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

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 assessements?
- How to model concept properties correctly such that annotations can be automatized?
- Do models generalize?
  Are users happy?

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 assessements?
- How to model concept properties correctly such that annotations can be automatized?
- Do models generalize? Are users happy?

Annotation Challenges

Questions

Annotation Challenges

Questions

- Is the task objectively decidable?
  (entities vs. entailment or translation)

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)

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)

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

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?

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?

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

Modeling Approaches

- Rule-based methods, lexicon-based approaches

Modeling Approaches

- Rule-based methods, lexicon-based approaches
  - + Transparent
  - + Can be well grounded in theories

Modeling Approaches

- Rule-based methods, lexicon-based approaches
  - + Transparent
  - + Can be well grounded in theories
  - – Often conceptually too simple

Modeling Approaches

- Rule-based methods, lexicon-based approaches
  - \+ Transparent
  - \+ Can be well grounded in theories
  - – Often conceptually too simple
  - – Difficult to achieve good performance

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

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

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

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

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

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

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

Prompting with Instruction-tuned

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

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!

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

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.

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

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!

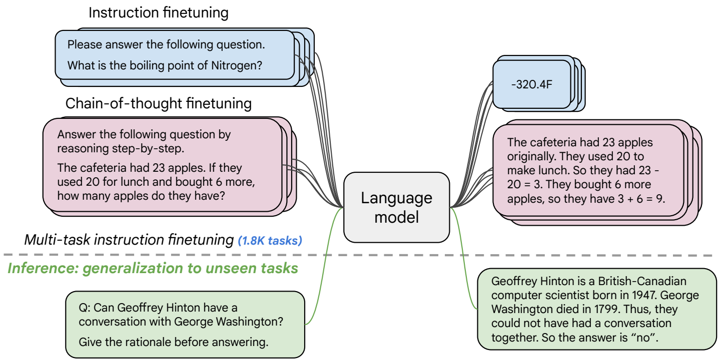
Step 2: Fine-tune this model to solve instructions

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

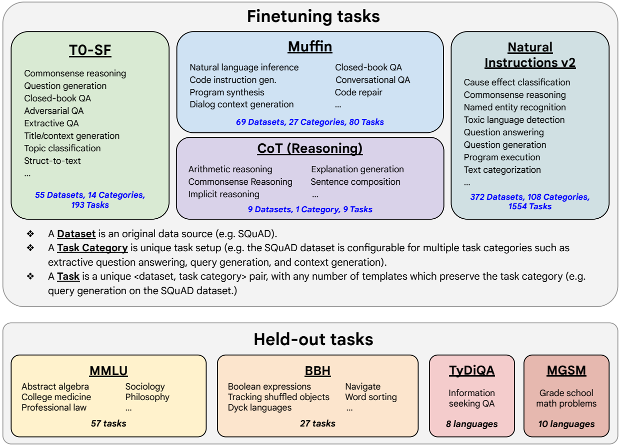
Step 3: Fine-tune with reinforcement learning from human-feedback on unseen tasks

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

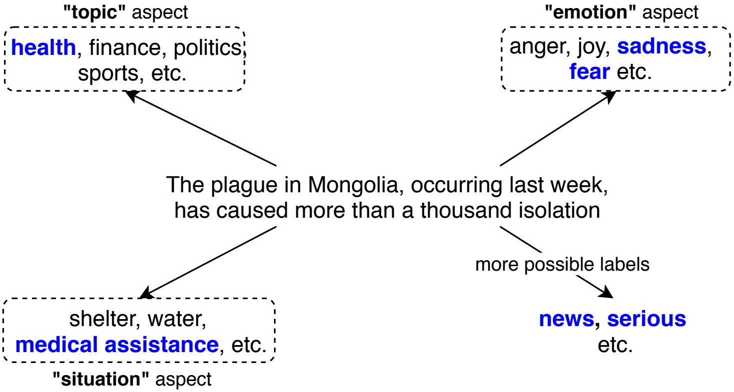
Example: Flan-T5 (1)



Example: Flan-T5 (2)



Text Classification as Natural Language Inference



W. Yin et al. (2019). "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach". In: *EMNLP-IJCNLP*

Observation

All three model types can be used for zero-shot text analysis

- Models that predict the next word or a missing word
- Models tuned for natural language inference
- Instruction-tuned models

Observation

All three model types can be used for zero-shot text analysis

- Models that predict the next word or a missing word
  - " 'He is happy.' The sentiment polarity of this statement is"
- Models tuned for natural language inference
- Instruction-tuned models

Observation

All three model types can be used for zero-shot text analysis

- Models that predict the next word or a missing word
  - " 'He is happy.' The sentiment polarity of this statement is"
- Models tuned for natural language inference
  - "He is happy" – "This sentence is positive."
- Instruction-tuned models

Observation

All three model types can be used for zero-shot text analysis

- Models that predict the next word or a missing word
  - " 'He is happy.' The sentiment polarity of this statement is"
- Models tuned for natural language inference
  - "He is happy" – "This sentence is positive."
- Instruction-tuned models
  - "What is the sentiment of the following sentence 'He is happy'? Answer with a digit only where 1 is positive and 2 is negative."

What's next?

What's next?

- We have now seen a set of methods to solve NLP tasks.

What's next?

- We have now seen a set of methods to solve NLP tasks.
  - NLI-based Zero-Shot Predictions
  - Fine-tuning language models; traditional ML/DL
  - Prompting with Instruction-tuned models

What's next?

- We have now seen a set of methods to solve NLP tasks.
  - NLI-based Zero-Shot Predictions
  - Fine-tuning language models; traditional ML/DL
  - Prompting with Instruction-tuned models
- I will now introduce emotion analysis.

What's next?

- We have now seen a set of methods to solve NLP tasks.
  - NLI-based Zero-Shot Predictions
  - Fine-tuning language models; traditional ML/DL
  - Prompting with Instruction-tuned models
- I will now introduce emotion analysis.
- Then I will show three examples from this area with different methods:

What's next?

- We have now seen a set of methods to solve NLP tasks.
  - NLI-based Zero-Shot Predictions
  - Fine-tuning language models; traditional ML/DL
  - Prompting with Instruction-tuned models
- I will now introduce emotion analysis.
- Then I will show three examples from this area with different methods:
  - NLI-based zero-shot emotion classification

What's next?

- We have now seen a set of methods to solve NLP tasks.
  - NLI-based Zero-Shot Predictions
  - Fine-tuning language models; traditional ML/DL
  - Prompting with Instruction-tuned models
- I will now introduce emotion analysis.
- Then I will show three examples from this area with different methods:
  - NLI-based zero-shot emotion classification
  - Traditional ML for appraisal-based corpus creation

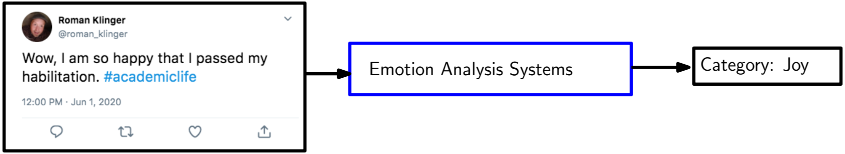
What's next?

- We have now seen a set of methods to solve NLP tasks.
  - NLI-based Zero-Shot Predictions
  - Fine-tuning language models; traditional ML/DL
  - Prompting with Instruction-tuned models
- I will now introduce emotion analysis.
- Then I will show three examples from this area with different methods:
  - NLI-based zero-shot emotion classification
  - Traditional ML for appraisal-based corpus creation
  - Prompt-based affective text generation

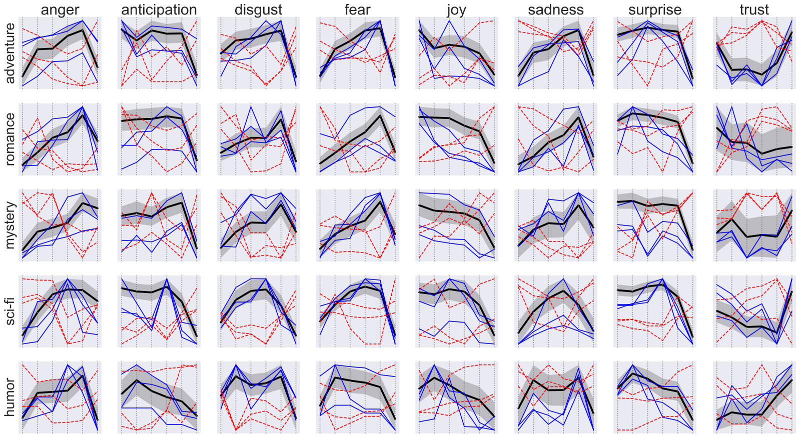
Outline

Emotion Analysis: What we want to do.

Emotion Analysis: What we want to do.



Literary Studies



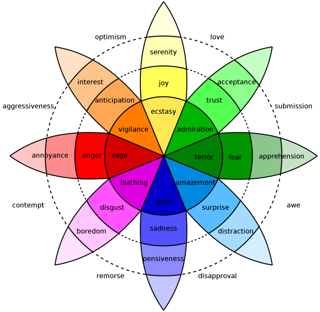
Kim et al., 2017.
Investigating the Relationship between Literary Genres and Emotional Plot Development. LaTeCH@ACL

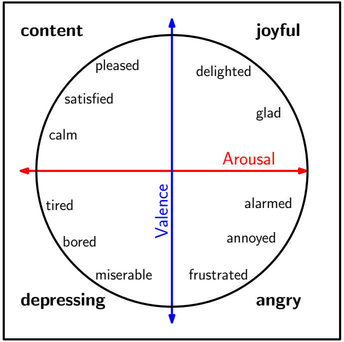
Dominant Emotions Expressed in News Articles

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?



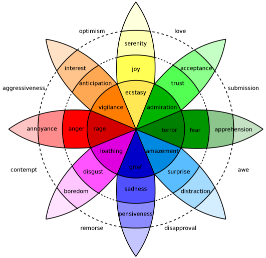


- Emotion models in psychology explain how emotions are developed.
- Text analysis models learn to associate textual realizations to emotion concepts. They do not (explicitly?) use knowledge from such theories.

Outline

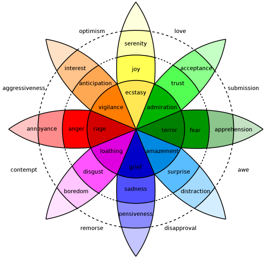
1. NLP Research Methods
2. Emotion Analysis
3. Zero-Shot Learning for Emotion Classification
4. Appraisal-based Emotion Analysis
5. Prompt Search for Text Generation
6. Take Home

ZSL for Emotion Classification



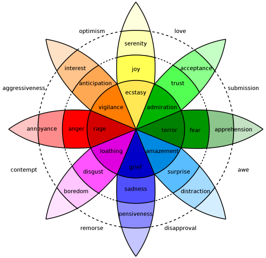
- Often used labels in emotion classification in text: Ekman's basic emotions or a subset from Plutchik's wheel

ZSL for Emotion Classification



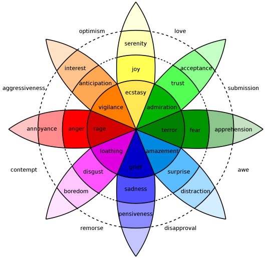
- Often used labels in emotion classification in text: Ekman's basic emotions or a subset from Plutchik's wheel
- Sometimes domains require specific sets

ZSL for Emotion Classification



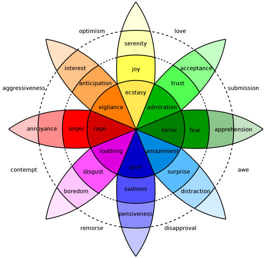
- Often used labels in emotion classification in text: Ekman's basic emotions or a subset from Plutchik's wheel
- Sometimes domains require specific sets
- "Joy, insecurity, annoyance, relaxation, and boredom" to model emotions of drivers (Cevher, Zepf, Klinger, KONVENS 2019)

ZSL for Emotion Classification



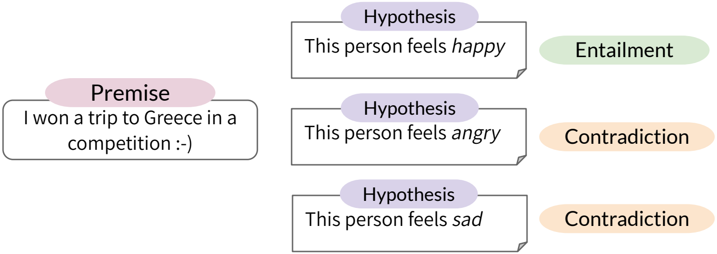
- Often used labels in emotion classification in text: Ekman's basic emotions or a subset from Plutchik's wheel
- Sometimes domains require specific sets
- "Joy, insecurity, annoyance, relaxation, and boredom" to model emotions of drivers (Cevher, Zepf, Klinger, KONVENS 2019)
- "Aesthetic emotions" for poetry (beauty, awe, suspense, uneasiness, sadness, …) (Haider, Eger, Kim, Klinger, Menninghaus, LREC 2020)

ZSL for Emotion Classification



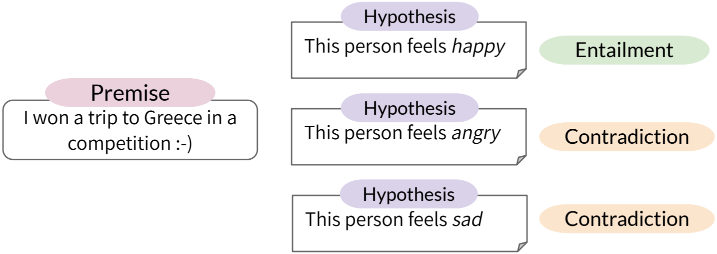
- Often used labels in emotion classification in text: Ekman's basic emotions or a subset from Plutchik's wheel
- Sometimes domains require specific sets
- "Joy, insecurity, annoyance, relaxation, and boredom" to model emotions of drivers (Cevher, Zepf, Klinger, KONVENS 2019)
- "Aesthetic emotions" for poetry (beauty, awe, suspense, uneasiness, sadness, ...) (Haider, Eger, Kim, Klinger, Menninghaus, LREC 2020)
- Do we need to create an emotion corpus with domain specific labels for every new application domain where the label set changes?

Emotion ZSL as Natural Language Inference



F. M. Plaza-del Arco et al. (2022). "Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora". In: *COLING*

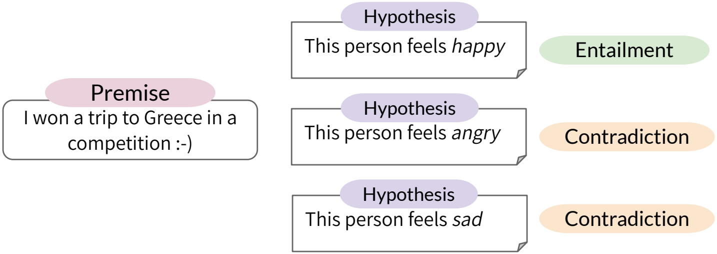
Emotion ZSL as Natural Language Inference



- Does it matter which NLI model we use as a backbone?

F. M. Plaza-del Arco et al. (2022). "Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora". In: *COLING*

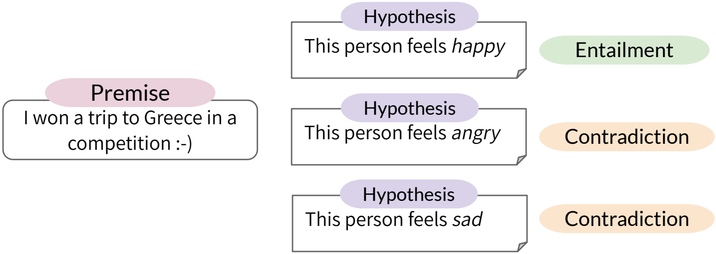
Emotion ZSL as Natural Language Inference



- Does it matter which NLI model we use as a backbone?
- How to represent the emotion?

F. M. Plaza-del Arco et al. (2022). "Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora". In: *COLING*

Emotion ZSL as Natural Language Inference



- Does it matter which NLI model we use as a backbone?
- How to represent the emotion?
- Does the hypothesis formulation need to be specific for a particular domain?

F. M. Plaza-del Arco et al. (2022). "Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora". In: *COLING*

Emotion Hypotheses

Emo-Name

Expr-Emo

Feels-Emo

WN-Def

Emo-S

Expr.-S

Feels.-S

Emotion Hypotheses

Emo-Name

angry

Expr-Emo

Feels-Emo

WN-Def

Emo-S

Expr.-S

Feels.-S

Emotion Hypotheses

Emo-Name	angry	**Emo-S**
Expr-Emo	This text expresses anger	**Expr.-S**
Feels-Emo		**Feels.-S**
WN-Def		

Emotion Hypotheses

Emo-Name
angry

Expr-Emo
This text expresses anger

Feels-Emo
This person feels anger

WN-Def

Emo-S

Expr.-S

Feels.-S

Emotion Hypotheses

Emo-Name
angry

Expr-Emo
This text expresses anger

Feels-Emo
This person feels anger

WN-Def

Emo-S

Expr.-S

Feels.-S

Same prefix + anger, annoyance, rage, outrage, fury, irritation

Emotion Hypotheses

Emo-Name	angry
Expr-Emo	This text expresses anger
Feels-Emo	This person feels anger
WN-Def	This person expresses a strong emotion; a feeling that is oriented toward some real or supposed grievance

Emo-S

Expr.-S

Feels.-S

Same prefix + anger, annoyance, rage, outrage, fury, irritation

Pretrained NLI Models

Data Set for Pretraining: MultiNLI Corpus, 433k sentence pairs:

Examples

Premise	**Label**	**Hypothesis**
Fiction		
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

- https://cims.nyu.edu/~sbowman/multinli/
- Pretrained models from Huggingface, we use RoBERTa, BART, and DeBERTa

Data Sets

Data Sets

- TEC (Mohammad 2012):
  Twitter corpus, automatically labeled with emotion hashtags
  Be the greatest dancer of your life! practice daily positive habits. [JOY]

Data Sets

- TEC (Mohammad 2012):
  Twitter corpus, automatically labeled with emotion hashtags
  Be the greatest dancer of your life! practice daily positive habits. [JOY]
- ISEAR (Scherer 1997):
  Descriptions of emotional events, triggered by emotion name
  When I was involved in a traffic accident. [FEAR]

Data Sets

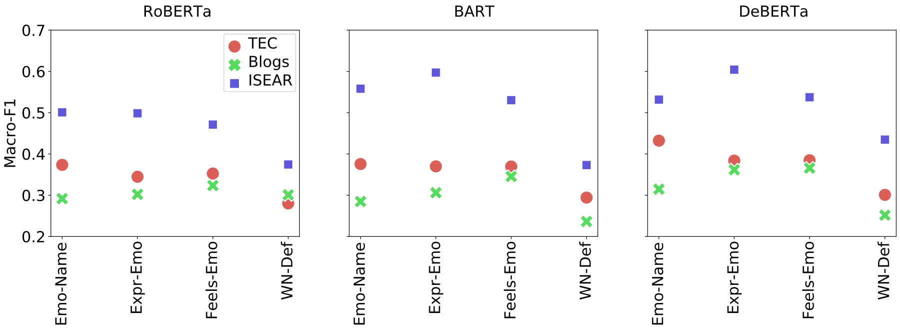
- TEC (Mohammad 2012):
  Twitter corpus, automatically labeled with emotion hashtags
  Be the greatest dancer of your life! practice daily positive habits. [JOY]
- ISEAR (Scherer 1997):
  Descriptions of emotional events, triggered by emotion name
  When I was involved in a traffic accident. [FEAR]
- Blogs (Aman 2007):
  Crowdsourced annotations of sentences from blogs
  I've never missed anyone so much as you. [SADNESS]

Data Sets

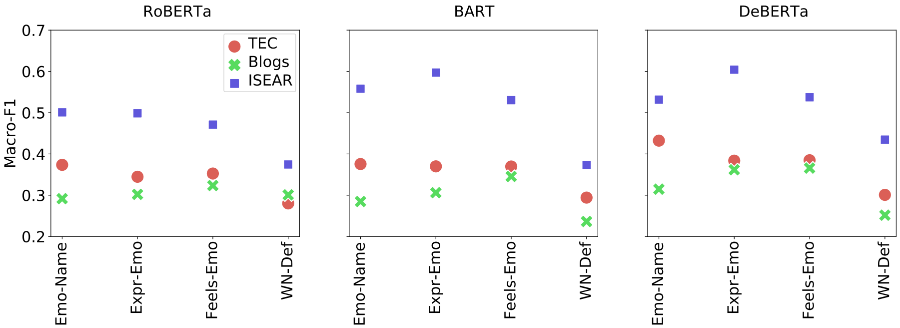
- TEC (Mohammad 2012):
  Twitter corpus, automatically labeled with emotion hashtags
  Be the greatest dancer of your life! practice daily positive habits. [JOY]
- ISEAR (Scherer 1997):
  Descriptions of emotional events, triggered by emotion name
  When I was involved in a traffic accident. [FEAR]
- Blogs (Aman 2007):
  Crowdsourced annotations of sentences from blogs
  I've never missed anyone so much as you. [SADNESS]

Emotion labels: anger, fear, joy, sadness, disgust, surprise, guilt, shame

The role of the NLI model

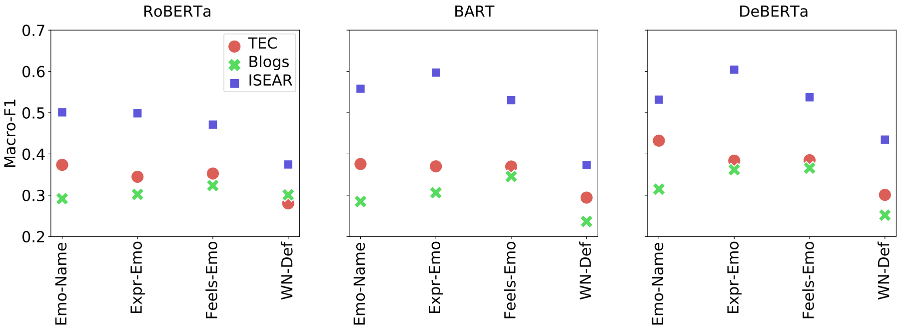


The role of the NLI model



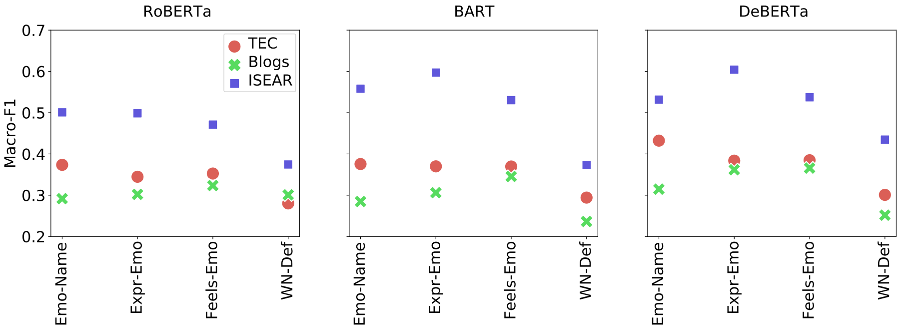
- Does the choice of the NLI model matter?

The role of the NLI model



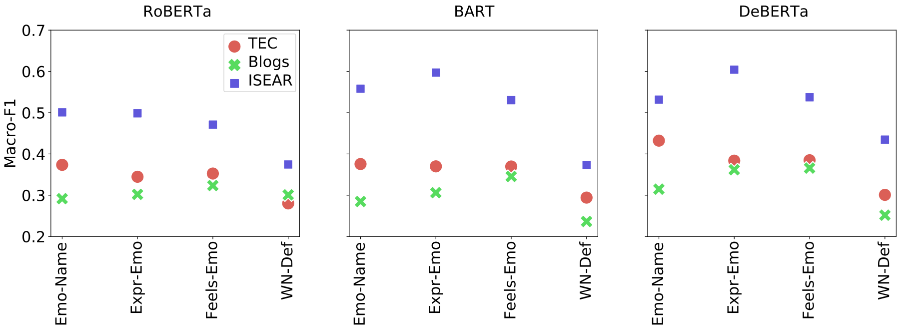
- Does the choice of the NLI model matter?
    - Performance differences between data sets are (mostly) independent of model

The role of the NLI model



- Does the choice of the NLI model matter?
  - Performance differences between data sets are (mostly) independent of model
- Does the prompt matter regarding the data set?

The role of the NLI model



- Does the choice of the NLI model matter?
  - Performance differences between data sets are (mostly) independent of model
- Does the prompt matter regarding the data set?
  - WN-Def always lowest performance

The role of the NLI model

- If one NLI model performs ZSL well on some domains, it also does so on others.

The role of the NLI model

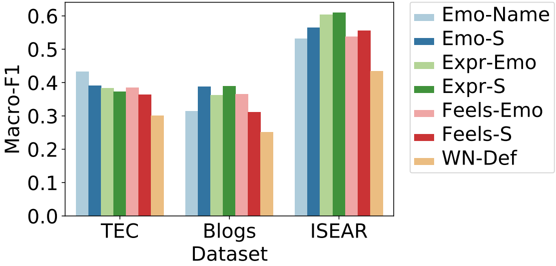
- If one NLI model performs ZSL well on some domains, it also does so on others.
- That's great! New, better models probably improve the results across domains.

The role of the prompt design

Is the emotion representation in the prompt specific to a domain/dataset?

The role of the prompt design

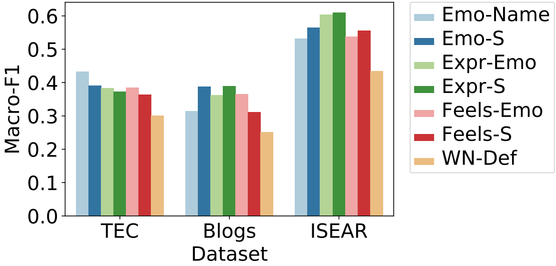
Is the emotion representation in the prompt specific to a domain/dataset?



- TEC: single emotion names work better than with synonyms

The role of the prompt design

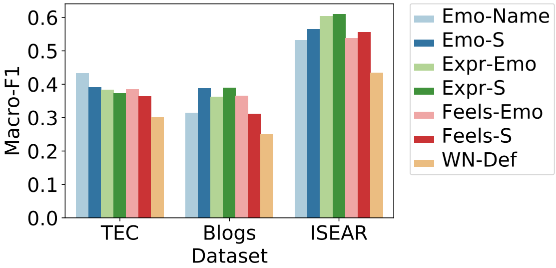
Is the emotion representation in the prompt specific to a domain/dataset?



- TEC: single emotion names work better than with synonyms
- BLOGS: synonyms harm the performance for Feels-Emo/S prompts

The role of the prompt design

Is the emotion representation in the prompt specific to a domain/dataset?



- TEC: single emotion names work better than with synonyms
- BLOGS: synonyms harm the performance for Feels-Emo/S prompts
- Generally: synonyms help, except for some cases, in which annotaton procedure might be the reason

The role of the prompt design (3)

There is not a single prompt which works well across all domains.

The role of the prompt design (3)

There is not a single prompt which works well across all domains.

But: Putting multiple prompts together in a model ensemble works nearly en par with individual single prompts.

Conclusion

Conclusion

- We showed the first evaluation of prompts across domains for emotion ZSL classification.

Conclusion

- We showed the first evaluation of prompts across domains for emotion ZSL classification.
- The concrete NLI model which forms the backbone seems not to matter.

Conclusion

- We showed the first evaluation of prompts across domains for emotion ZSL classification.
- The concrete NLI model which forms the backbone seems not to matter.
- There is not one individual prompt which works best for each domain

Where are we?

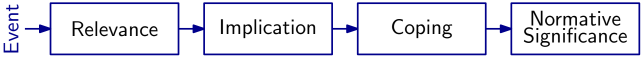
- We wanted to achieve a domain-independent and label-set independent model.
- We did pretty much achieve this, but the performance is lower than traditional machine learning methods.

Outline

Appraisal Theories

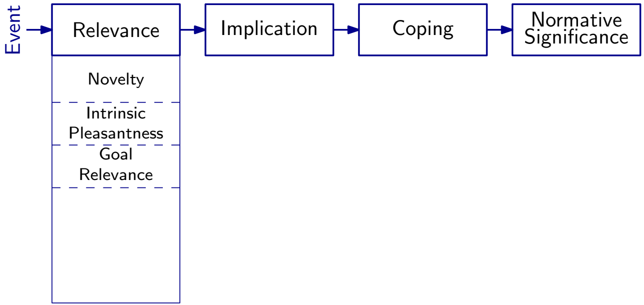
- Appraisal theories explain the relation between emotions based on other dimensions.
- If we can build appraisal predictors, this might help to have more robust emotion prediction models.

Cognitive Appraisal in Scherer's Component Process model



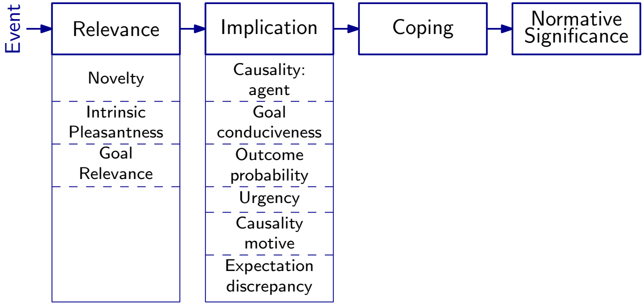
K.R. Scherer (2001). Appraisal Considered as a Process of Multilevel Sequential Checking.

Cognitive Appraisal in Scherer's Component Process model



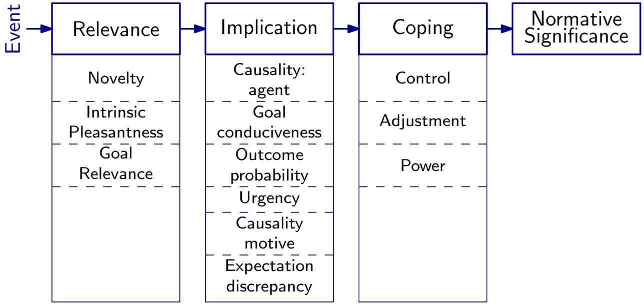
K.R. Scherer (2001). Appraisal Considered as a Process of Multilevel Sequential Checking.

Cognitive Appraisal in Scherer's Component Process model



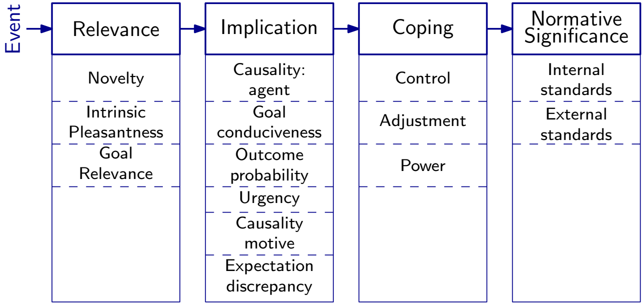
K.R. Scherer (2001). Appraisal Considered as a Process of Multilevel Sequential Checking.

Cognitive Appraisal in Scherer's Component Process model



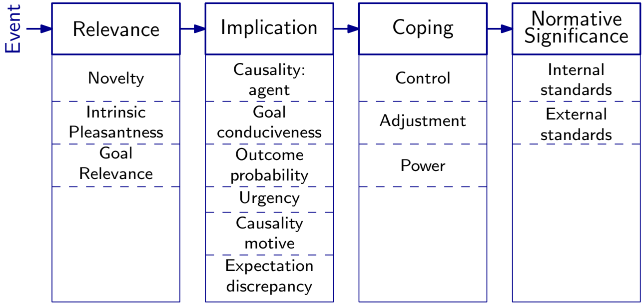
K.R. Scherer (2001). Appraisal Considered as a Process of Multilevel Sequential Checking.

Cognitive Appraisal in Scherer's Component Process model



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

Cognitive Appraisal in Scherer's Component Process model



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

Perhaps appraisals are an alternative, more general approach to emotion analysis?

Research Questions

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*

Research Questions

- Can appraisals be annotated reliably?

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*

Research Questions

- Can appraisals be annotated reliably?
- Can we predict appraisal variables from event descriptions?

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*

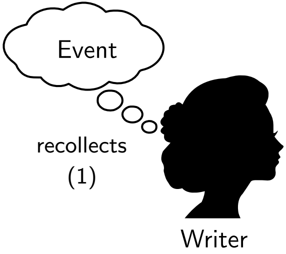
Research Questions

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

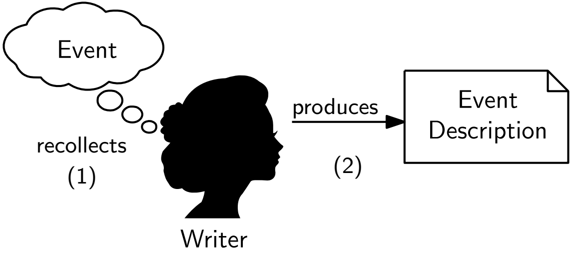
E. Troiano et al. (2023). "Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction". In: *Computational Linguistics* 49.1

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

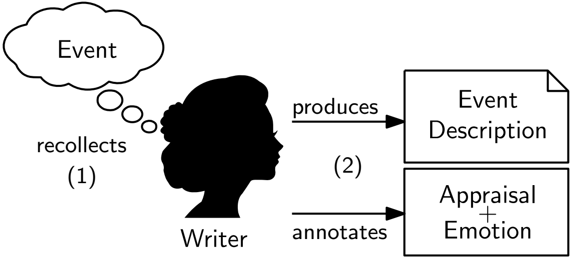
Approach



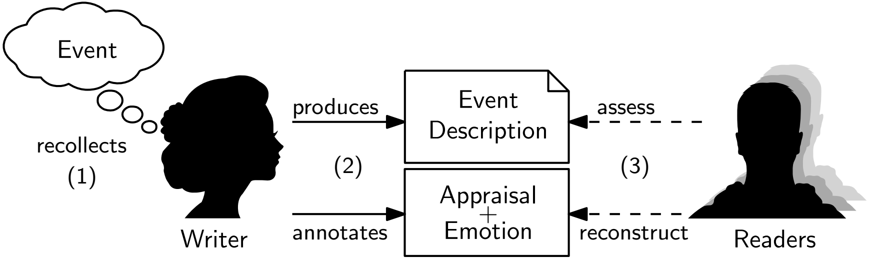
Approach



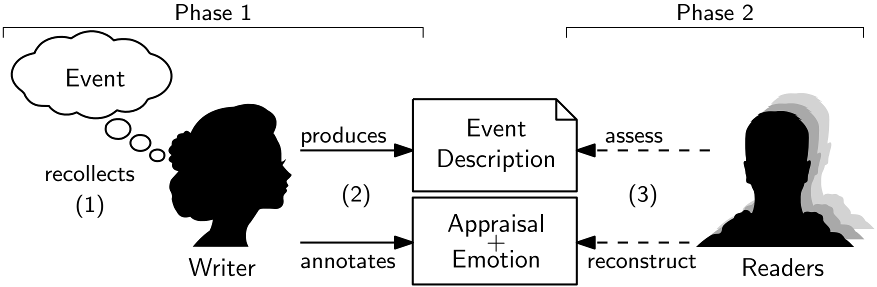
Approach


Event
recollects
(1)
Writer
produces
(2)
annotates
Event
Description
Appraisal
+
Emotion

Approach

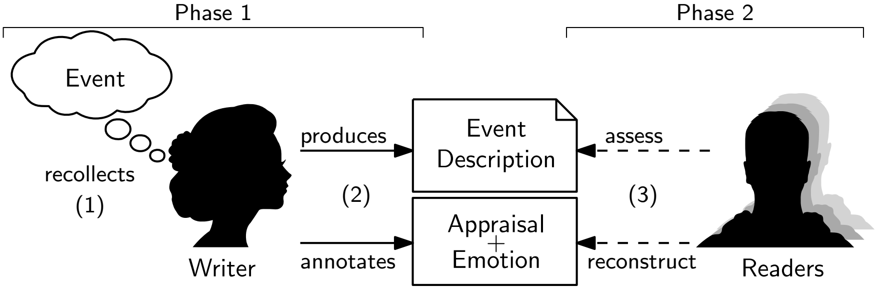

Event
recollects
(1)
Writer
produces
(2)
annotates
Event
Description
Appraisal
+
Emotion
assess
(3)
reconstruct
Readers

Approach



Phase 1
Phase 2
Event
recollects
(1)
Writer
produces
annotates
(2)
Event Description
Appraisal + Emotion
assess
reconstruct
(3)
Readers

Approach



- Production: 550 event descriptions for anger, boredom, disgust, fear, guilt/shame, joy, pride, relief, sadness, surprise, trust, no emotion

Examples

Examples

pride I baked a delicious strawberry cobbler.

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.

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

Setup:

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

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)

Setup:

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

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.

Setup:

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

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?

Setup:

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

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.

Setup:

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

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.

Setup:

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

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?

Setup:

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

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)

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

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

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

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

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

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

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

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

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

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

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

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

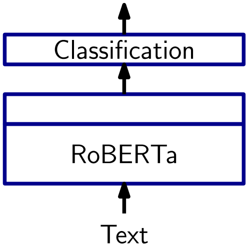
Examples (writer/reader/avg. writer–reader agreement as error)

- 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
  shame, relief, 1.8 I tasked with sorting out some files from the office the previous day and I slept off when I got home

Modeling Results

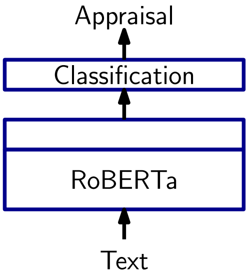
Modeling Results

- Classification with RoBERTa-based models



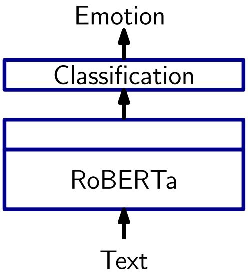
Modeling Results

- Classification with RoBERTa-based models
- Appraisal Classification: 75 F_1



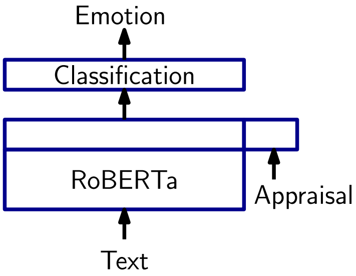
Modeling Results

- Classification with RoBERTa-based models
- Appraisal Classification: 75 F_1
- Emotion classification: 59 F_1



Modeling Results

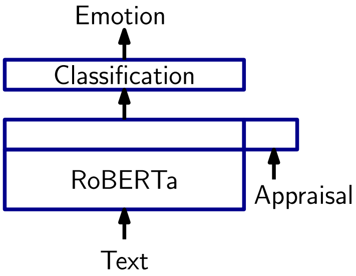
- Classification with RoBERTa-based models
- Appraisal Classification: 75 F_1
- Emotion classification: 59 F_1
- + Appraisals: +2pp F_1
  (+10 for guilt, +6 for sadness)



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)

$\Rightarrow$ Appraisals help to build better models.



Examples where Appraisals correct the Emotion Classifier

Examples where Appraisals correct the Emotion Classifier

- When my child settled well into school

trust→relief

Examples where Appraisals correct the Emotion Classifier

- When my child settled well into school

  trust→relief

- broke an expensive item in a shop accidently

  guilt→shame

Examples where Appraisals correct the Emotion Classifier

- When my child settled well into school

  trust→relief

- broke an expensive item in a shop accidently

  guilt→shame

- my mother made me feel like a child

  shame→anger

Examples where Appraisals correct the Emotion Classifier

- When my child settled well into school

  trust→relief
- broke an expensive item in a shop accidently

  guilt→shame
- my mother made me feel like a child

  shame→anger
- I passed my Irish language test

  pride→relief

Examples where Appraisals correct the Emotion Classifier

- When my child settled well into school
  trust→relief
- broke an expensive item in a shop accidently
  guilt→shame
- my mother made me feel like a child
  shame→anger
- I passed my Irish language test
  pride→relief
- His toenails where massive
  pride→disgust

Conclusion & Summary

Conclusion & Summary

- We presented the first self-annotated large-scale appraisal corpus

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)

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

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

Conditional Text Generation via Prompt Search

Conditional Text Generation via Prompt Search

Task

Generate text that expresses a predefined emotion.

Conditional Text Generation via Prompt Search

Task

Generate text that expresses a predefined emotion.

Applications

Module of dialogue systems, intelligent agents, or (story-)writing assistants.

Conditional Text Generation via Prompt Search

Task

Generate text that expresses a predefined emotion.

Applications

Module of dialogue systems, intelligent agents, or (story-)writing assistants.

State of the art

Guide the decoding of the text generation with emotion specific information.

Conditional Text Generation via Prompt Search

Task

Generate text that expresses a predefined emotion.

Applications

Module of dialogue systems, intelligent agents, or (story-)writing assistants.

State of the art

Guide the decoding of the text generation with emotion specific information.

Novel Method

Automatic prompt optimization of instruction-tuned models.

Method: Automatic Prompt Optimization

Method: Automatic Prompt Optimization

- Seed prompt: "Write a text that expresses ⟨emotion⟩"

Method: Automatic Prompt Optimization

- Seed prompt: "Write a text that expresses ⟨emotion⟩"
- Paraphrase: Replace/add/remove words in seed prompts

Method: Automatic Prompt Optimization

- Seed prompt: "Write a text that expresses ⟨emotion⟩"
- Paraphrase: Replace/add/remove words in seed prompts
- Evaluation: Run generation process, evaluate with pre-trained emotion classifier

Method: Automatic Prompt Optimization

- Seed prompt: "Write a text that expresses ⟨emotion⟩"
- Paraphrase: Replace/add/remove words in seed prompts
- Evaluation: Run generation process, evaluate with pre-trained emotion classifier
- Selection: Keep paraphrased prompts if performance is improved

Example of Input/Output Pairs during Optimization (Flan-T5)

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don’t know what to do about it.
- Write a text to expresses disgust

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame
  - i don't know how to express my shame.

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame
  - i don't know how to express my shame.
- Write a text that to expresses shame

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame
  - i don't know how to express my shame.
- Write a text that to expresses shame
  - i'm sorry to hear that.

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame
  - i don't know how to express my shame.
- Write a text that to expresses shame
  - i'm sorry to hear that.
- Write in a long enough string to expresses joy

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame
  - i don't know how to express my shame.
- Write a text that to expresses shame
  - i'm sorry to hear that.
- Write in a long enough string to expresses joy
  - a long enough string to express joy.

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame
  - i don't know how to express my shame.
- Write a text that to expresses shame
  - i'm sorry to hear that.
- Write in a long enough string to expresses joy
  - a long enough string to express joy.
- Write a long text string to expresses joy

Example of Input/Output Pairs during Optimization (Flan-T5)

- Write a text that expresses disgust
  - i don't know what to do about it.
- Write a text to expresses disgust
  - i'm disgusted.
- a text that expresses fear
  - a text that expresses fear
- Write in a text to expresses shame
  - i don't know how to express my shame.
- Write a text that to expresses shame
  - i'm sorry to hear that.
- Write in a long enough string to expresses joy
  - a long enough string to express joy.
- Write a long text string to expresses joy
  - i love you so much

Take Home

Take Home

- NLP Research is driven by task definitions and appropriate modeling

Take Home

- NLP Research is driven by task definitions and appropriate modeling
- Natural language inference can be applied for emotion classification – without knowing the emotion categories in advance

Take Home

- NLP Research is driven by task definitions and appropriate modeling
- Natural language inference can be applied for emotion classification – without knowing the emotion categories in advance
- Annotation and language model fine-tuning: Appraisal theories as a novel approach to emotion analysis in text – they support emotion classification and also do not require to fully specify the emotion set

Take Home

- NLP Research is driven by task definitions and appropriate modeling
- Natural language inference can be applied for emotion classification – without knowing the emotion categories in advance
- Annotation and language model fine-tuning: Appraisal theories as a novel approach to emotion analysis in text – they support emotion classification and also do not require to fully specify the emotion set
- Automatic prompt optimization: Emotion-conditioned text generation – very challenging to automatically find well-performing prompts.

Thank you for
your attention.
Questions? Remarks?

?

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Thanks to

- Ph.D. Students
  - Enrica Troiano
  - Laura Oberländer née Bostan
  - Yarik Menchaca Resendiz
  - Flor Miriam Plaza Del Arco
- Collaborators
  - Kai Sassenberg

Universität Stuttgart
Institut für
Maschinelle Sprachverarbeitung

Natural Language Processing Tasks and Methods

Challenges for Emotion Analysis and Generation

ZPID, Dec 13, 2023

Roman Klinger
roman.klinger@ims.uni-stuttgart.de

@roman_klinger romanklinger
https://www.romanklinger.de/

https://www.romanklinger.de/talks/zpid2023.pdf

Appraisal Variables

Relevance	Implication	Coping	Normative Significance
Novelty	Causality: agent	Control	Internal standards compatibility
(1) suddenness	(7) own responsibility	(19) own control*	(14) clash with own standards/ideals
(2) familiarity	(8) other's respons.	(20) others' control*	
(3) predictability	(9) situational respons.	(21) chance control*	
(16) attention*			
(17) att. removal*	Goal conduciveness	Adjustment	External standards compatibility
	(10) goal support	(13) anticipated acceptance	(15) clash with laws/norms
Intrinsic Pleasantness			
(4) pleasant	Outcome probability	(18) effort*	
(5) unpleasant	(11) consequence anticipation		
Goal Relevance			
(6) goal-related	Urgency		
	(12) response urgency		

Variable Assessement

Appraisal Variables

(1) The event was sudden or abrupt. (*suddenness*)
(2) The event was familiar. (*familiarity*)
(3) I could have predicted the occurrence of the event. (*event predictability*)
(4) The event was pleasant. (*pleasantness*)
(5) The event was unpleasant. (*unpleasantness*)
(6) I expected the event to have important consequences for me. (*goal relevance*)
(7) The event was caused by my own behavior. (*own responsibility*)
(8) The event was caused by somebody else's behavior. (*other responsibility*)
(9) The event was caused by chance, special circumstances, or natural forces. (*situational responsibility*)
(10) I expected positive consequences for me. (*goal support*)
(11) I anticipated the consequences of the event. (*anticip. conseq.*)
(12) The event required an immediate response. (*urgency*)
(13) I anticipated that I would easily live with the unavoidable consequences of the event. (*accept. conseq.*)
(14) The event clashed with my standards and ideals. (*internal standards*)
(15) The actions that produced the event violated laws or socially accepted norms. (*external norms*)
(16) I had to pay attention to the situation. (*attention*)
(17) I tried to shut the situation out of my mind. (*not consider*)
(18) The situation required me a great deal of energy to deal with it. (*effort*)
(19) I was able to influence what was going on during the event. (*own control*)
(20) Someone other than me was influencing what was going on. (*others' control*)
(21) The situation was the result of outside influences of which nobody had control. (*situational control*)

- All variables are similarly assessed by writers and readers

Additional Variables

- Age, Gender
- Ethnicity, Education
- Event familiarity
  for readers
- Personality traits (Gosling 2003)
  - openness
  - conscientiousness
  - extraversion
  - agreeablenes
  - emotional stability

Reliability Results

			Agreement					
			Emotion				Appraisal	
			F_1		Acc.		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

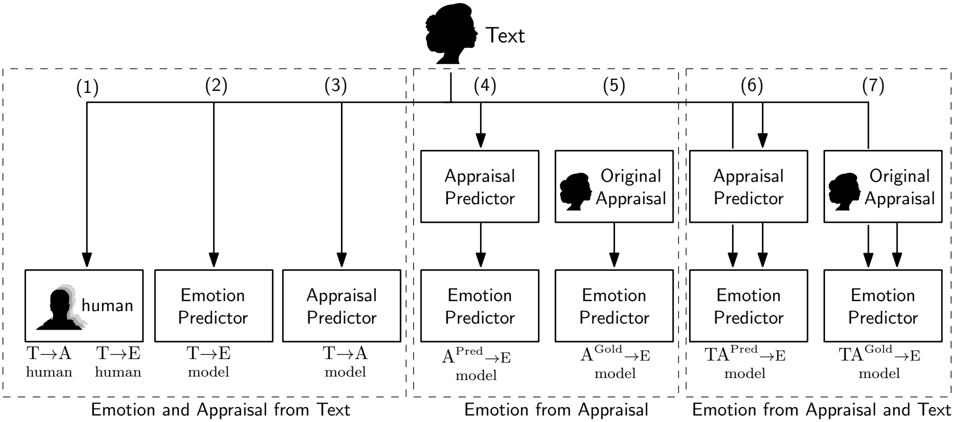
- Validators agree more with each other than with the generator
- Females agree more with each other on emotions
- Validators agree more if their age difference is small
- Event familiarity hurts agreement for appraisal
- We expected Open annotators to perform better.
- Emotional stability "hurts" emotion annotation.
- Extraversion, Conscient., Agreeableness help.
- Most differences are quite small (though significant)

Appraisals add additional information to emotion analysis

That I put together a funeral service for my Aunt

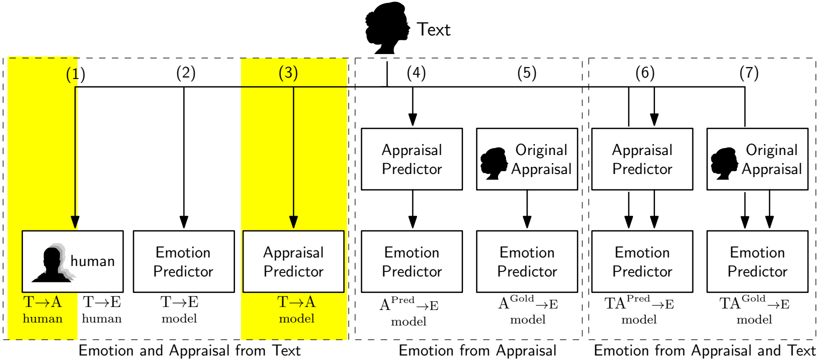
Dimension	Writer	Readers	Δ
Emotion	Pride	Sadness	
Suddenness	4	3.6	0.4
Familiarity	1	2.0	−1.0
Predictability	1	1.8	−0.8
Pleasantness	4	1.0	3.0
Unpleasantness	2	4.8	−2.8
Goal-Relevance	4	2.6	1.4
Chance-Resp.	4	4.4	−0.4
Self-Resp.	1	1.2	−0.2
Other-Resp.	1	1.4	−0.4
Conseq.-Predict.	2	1.8	0.2
Goal Support	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

Setup



- All models are RoBERTa-based, additional information is added to the penultimate layer before the output

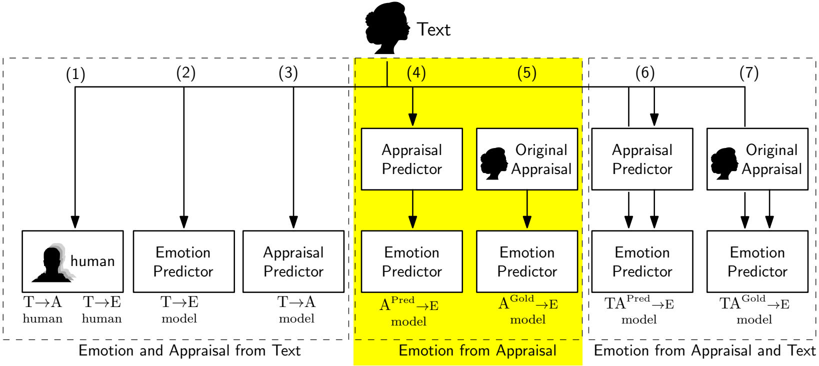
How well can we predict appraisals?



⇒ 75 macro F_1 for model

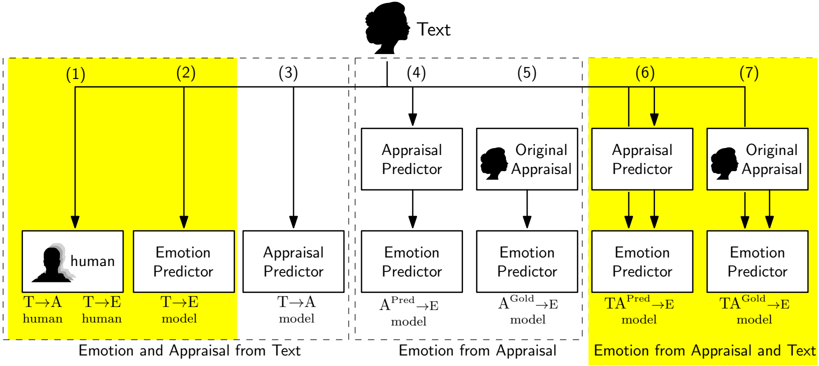
- Model outperforms humans by 13pp F_1
- Slightly unfair assessement:
  Model gets to see generator's labels during training, humans do not

Are appraisals sufficient to infer the emotion?



- Skipping this today
- Short answer: comparably (low) to psychological studies

Do appraisals help emotion prediction?



Results in F_1

Emotion	(a) T→E human	(b) T→E model	$\Delta^{(b)}_{(a)}$	(c) TA^{Gold}→E model	$\Delta^{(c)}_{(b)}$	(d) TA^{Pred}→E model	$\Delta^{(d)}_{(c)}$	$\Delta^{(d)}_{(b)}$
Anger	57	53	−4	57	+4	57	0	+4
Boredom	73	84	+11	83	−1	83	0	−1
Disgust	65	66	+1	66	0	66	0	0
Fear	73	65	−8	67	+2	67	0	+2
Guilt	53	48	−5	58	+10	56	−2	+8
Joy	49	45	−4	48	+3	47	−1	+2
No-emotion	33	55	+22	56	+1	56	0	+1
Pride	59	54	−5	55	+1	55	0	+1
Relief	64	63	−1	62	−1	62	0	−1
Sadness	63	59	−4	65	+6	63	−2	+4
Shame	48	51	+3	50	−1	49	−1	−2
Surprise	42	53	+11	49	−4	50	+1	−3
Trust	52	74	+22	73	−1	72	−1	−2
Macro avg.	56	59	+3	61	+2	60	−1	+1

- On average, the model is as good as humans, but differs across emotions
- Adding gold appraisal helps some emotions considerably
- Predicted appraisals are similarly valuable as gold appraisals