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/
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?

- 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 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: Machine Translation de-en

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

- 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.
We finetune both with and without exemplars (i.e., zero-shot and few-shot) and with and without chain-of-thought, enabling generalization across a range of evaluation scenarios.

Example: Flan-T5 (1)
Example: Flan-T5 (2)

Finetuning tasks

TO-SF
- Commonsense reasoning
- Question generation
- Closed-book QA
- Adversarial QA
- Extractive QA
- Title/context generation
- Struct-to-text

Muffin
- Natural language inference
- Code instruction gen.
- Program synthesis
- Dialog context generation

CoT (Reasoning)
- Arithmetic reasoning
- Commonsense Reasoning
- Implicit reasoning

Natural Instructions v2
- Cause effect classification
- Commonsense reasoning
- Named entity recognition
- Toxic language detection
- Question answering
- Program execution
- Text categorization

Held-out tasks

MMLU
- Abstract algebra
- College medicine
- Professional law
- Sociology
- Philosophy

BBH
- Boolean expressions
- Tracking shuffled objects
- Dyck languages
- Navigate
- Word sorting

TyDiQA
- Information seeking QA

MGSM
- Grade school math problems

- 193 Tasks
- 69 Datasets, 27 Categories, 80 Tasks
- 55 Datasets, 14 Categories, 193 Tasks
- 9 Datasets, 1 Category, 9 Tasks
- 1554 Tasks

A Dataset is an original data source (e.g. SQuAD).
A Task Category is unique task setup (e.g. the SQuAD dataset is configurable for multiple task categories such as extractive question answering, query generation, and context generation).
A Task is a unique <dataset, task category> pair, with any number of templates which preserve the task category (e.g. query generation on the SQuAD dataset.)
Text Classification as Natural Language Inference

"topic" aspect

health, finance, politics, sports, etc.

"emotion" aspect

anger, joy, sadness, fear etc.

"situation" aspect

shelter, water, medical assistance, etc.

The plague in Mongolia, occurring last week, has caused more than a thousand isolation

more possible labels

news, serious etc.

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?

- 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

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
Emotion Analysis: What we want to do.

Wow, I am so happy that I passed my habilitation. #academiclife

12:00 PM · Jun 1, 2020

Emotion Analysis Systems

Category: Joy
Literary Studies

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

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Roman Klinger
ZPID, Dec 13, 2023
Dominant Emotions Expressed in News Articles

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Dominant Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>The Blaze, The Daily Wire, BuzzFeed</td>
</tr>
<tr>
<td>Annoyance</td>
<td>Vice, NewsBusters, AlterNet</td>
</tr>
<tr>
<td>Disgust</td>
<td>BuzzFeed, The Hill, NewsBusters</td>
</tr>
<tr>
<td>Fear</td>
<td>The Daily Mail, Los Angeles Times, BBC</td>
</tr>
<tr>
<td>Guilt</td>
<td>Fox News, The Daily Mail, Vice</td>
</tr>
<tr>
<td>Joy</td>
<td>Time, Positive.News, BBC</td>
</tr>
<tr>
<td>Pessimism</td>
<td>MotherJones, Intercept, Financial Times</td>
</tr>
<tr>
<td>Neg. Surprise</td>
<td>The Daily Mail, MarketWatch, Vice</td>
</tr>
<tr>
<td>Optimism</td>
<td>Bussines Insider, The Week, The Fiscal Times</td>
</tr>
<tr>
<td>Sadness</td>
<td>The Daily Mail, CNN, Daily Caller</td>
</tr>
<tr>
<td>Trust</td>
<td>The Daily Signal, Fox News, Mother Jones</td>
</tr>
</tbody>
</table>

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.
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?

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Roman Klinger
ZPID, Dec 13, 2023
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

<table>
<thead>
<tr>
<th>Emo-Name</th>
<th>angry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expr-Emo</td>
<td>This text expresses anger</td>
</tr>
<tr>
<td>Feels-Emo</td>
<td>This person feels anger</td>
</tr>
<tr>
<td>WN-Def</td>
<td>This person expresses a strong emotion; a feeling that is oriented toward some real or supposed grievance</td>
</tr>
</tbody>
</table>

| Emo-S | Same prefix + anger, |
| Expr.-S | annoyance, rage, outrage, fury, irritation |
| Feels.-S | |

Roman Klinger
ZPID, Dec 13, 2023
## Pretrained NLI Models

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

<table>
<thead>
<tr>
<th>Examples</th>
<th>Label</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fiction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Old One always comforted Ca’daan, except today.</td>
<td><em>neutral</em></td>
<td>Ca’daan knew the Old One very well.</td>
</tr>
<tr>
<td><strong>Letters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Your gift is appreciated by each and every student who will benefit from your generosity.</td>
<td><em>neutral</em></td>
<td>Hundreds of students will benefit from your generosity.</td>
</tr>
<tr>
<td><strong>Telephone Speech</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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</td>
<td><em>contradiction</em></td>
<td>August is a black out month for vacations in the company.</td>
</tr>
<tr>
<td><strong>9/11 Report</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At the other end of Pennsylvania Avenue, people began to line up for a White House tour.</td>
<td><em>entailment</em></td>
<td>People formed a line at the end of Pennsylvania Avenue.</td>
</tr>
</tbody>
</table>

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

- 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.
- 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?

- 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 annotation procedure might be the reason
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

• 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

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
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


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

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


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 \textit{(writer/reader/avg. writer–reader agreement as error)}

- All writers/readers agree on emotion, \textit{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, \textit{low} average appraisal agreement
  - disgust, 2.0
  - fear, 2.1

- All readers agree on the emotion, but \textit{not with the writer}, \textit{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 \textit{not with the writer}, \textit{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
  - trust → relief

- broke an expensive item in a shop accidentally
  - guilt → shame

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

- I passed my Irish language test
  - pride → relief

- His toenails were massive
  - 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. Zero-Shot Learning for Emotion Classification
4. Appraisal-based Emotion Analysis
5. Prompt Search for Text Generation
6. Take Home
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

- **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)

- 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

- 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?

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