Using Large Language Models to Detect Deliberative Elements in Public Discourse

DETECTING SUBJECTIVE EMOTIONS

Author: Bente Zuurbier Responsible professor: Luciano Cavalcante Siebert
Supervisors: Amir Homayounirad & Enrico Liscio

1. Introduction

Public discourse

- Allows people to express their opinions
- **Mediated discourse** helps people understand each other and change their point of view [1][2]
- Scaling up public discourse is difficult

Emotions:

- Detecting and handling **emotions** properly greatly helps a mediator [3][4]
- **Negative emotions:** participants distracted, manipulatable and irrational [3]
- **Positive emotions:** participants understanding and tell wants and needs [3]
- Studies found at least **27 distinct emotions** exist [5]
- **Emotion taxonomy** of 27 emotions and "neutral" created by GoEmotions [6]

Subjective labels:

- Emotions are highly **subjective**, no "true labels" exist [7]
- **Hard multi-label** and **soft labels** are used [7]

Large Language Models (LLMs) prompting strategies:

- **Zeroshot**: no examples are given alongside the prompt
- **Oneshot**: one example is given alongside the prompt
- **Fewshot**: small number of examples is given to the LLM to train on
- **Chain of thought**: LLM is asked to reason about intermediate steps

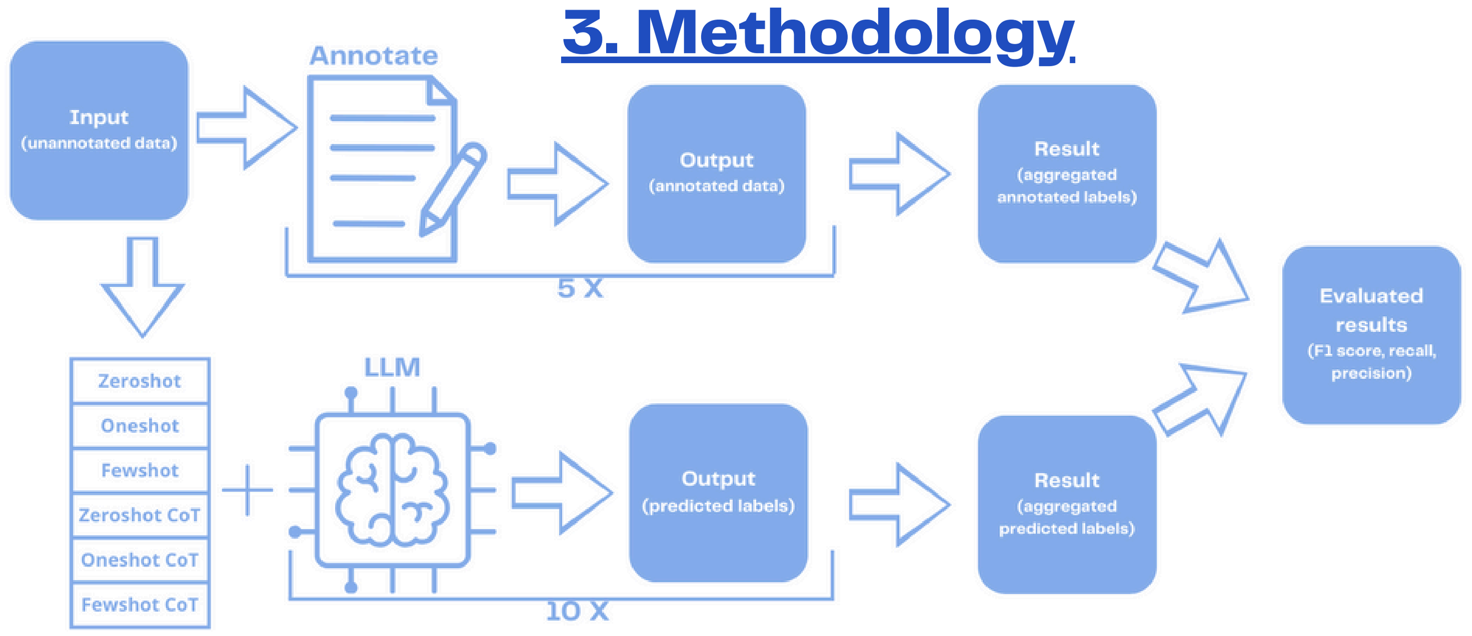
2. Research question

"How can Large Language Models be used to detect subjective emotions in public discourse?"

Sub-questions:

1. How can a LLM be **modelled** to detect subjective emotions in public discourse?
2. What is the effect of **different prompting stragies** on the accuracy of subjective emotion detection in Dutch public discourse by a LLM?
3. What is the effect of **different types of labels** on the accuracy of subjective emotion detection in Dutch public discourse?

3. Methodology



Annotate and aggregate data:

1. Annotator reads sentence and assigns emotion labels
2. Aggregate labels
3. Calculate the inter annotator agreement using Fleiss Kappa

Create the prompts:

1. Write prompt corresponding to the chosen method (zero-, one-, fewshot or chain of thought)

Experiment set-up:

1. Run LLM with prompt on the dataset 10x, aggregate labels
2. Evaluate the results

4. Results

Fleiss Kappa score: 0.00365

Hard Majority Labels

Training method	Micro F1 score	Recall	Precision
Zeroshot	0.385	0.420	0.355
Oneshot	0.469	0.580	0.394
Fewshot	0.486	0.537	0.444
Zeroshot Chain of Thought	0.410	0.399	0.422
Oneshot Chain of Thought	0.495	0.558	0.445
Fewshot Chain of Thought	0.480	0.485	0.474

Table 1: F1 score, recall and precision per prompting strategy

- **Annotated labels**: >=2 annotators picked it
- **Predicted labels:** >=2 LLM runs predicted it

Hard per LLM Run Labels

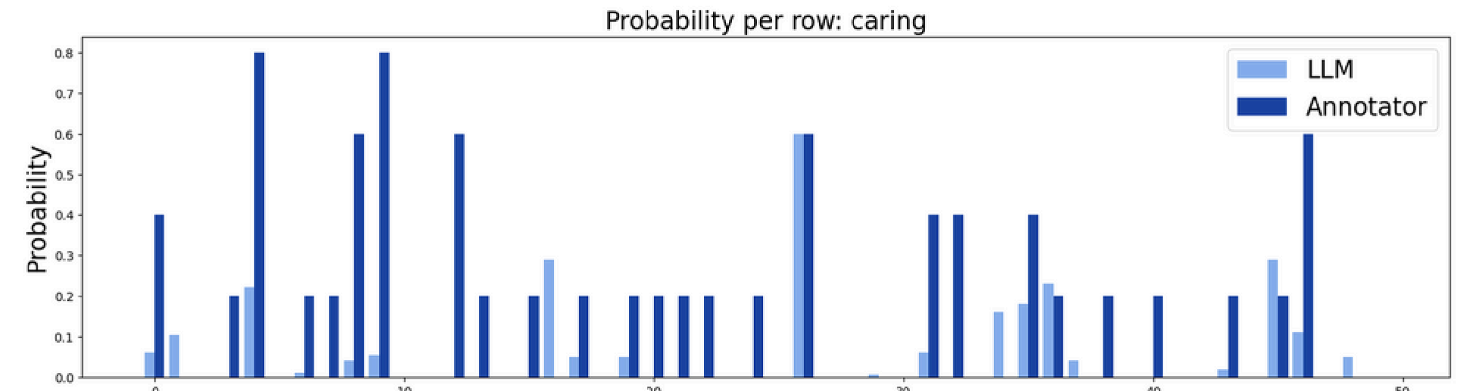
	Precision		Correct Labels		Incorrect Labels	
Training method	M	SD	M	SD	M	SD
Zeroshot	0,660	0,0248	54,4	3,720	28,0	2,145
Oneshot	0,720	0,0147	93,0	2,145	36,1	2,0
Fewshot	0,768	0,0304	75,8	2,857	23,0	3,0
Zeroshot Chain of Thought	0,709	0,0441	39,5	3,722	16,3	3,132
Oneshot Chain of Thought	0,718	0,0221	68,5	2,377	27,0	2,864
Fewshot Chain of Thought	0,764	0,0244	52,6	3,137	16,3	2,492

Table 2: Precision, correct and incorrect labels per prompting strategy

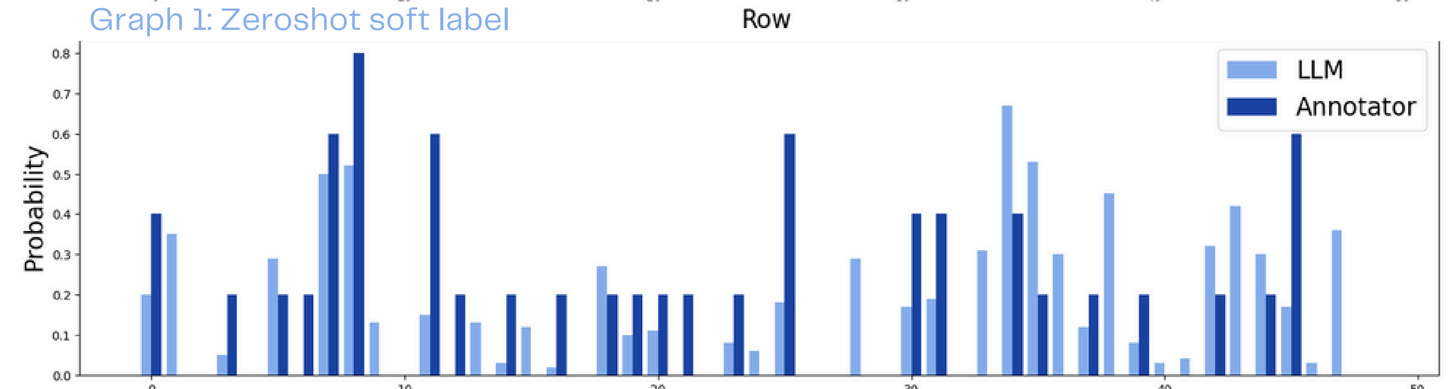
- **Annotated labels**: >=1 annotators picked it
- **Predicted labels:** per LLM run

Soft Labels

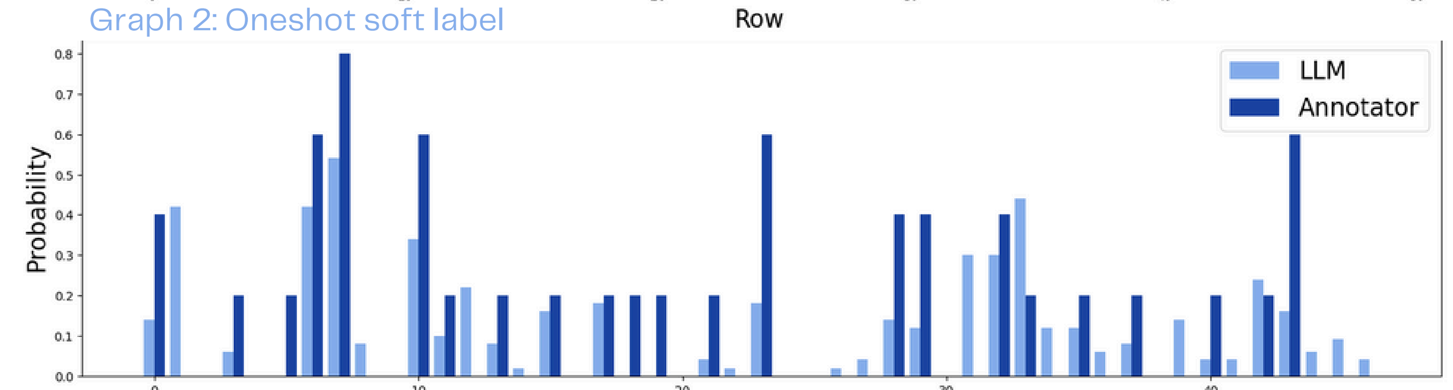
- **Annotated labels**: sum labels / number of annotators
- **Predicted labels:** sum labels / number of LLM runs



Graph 1: Zeroshot soft label

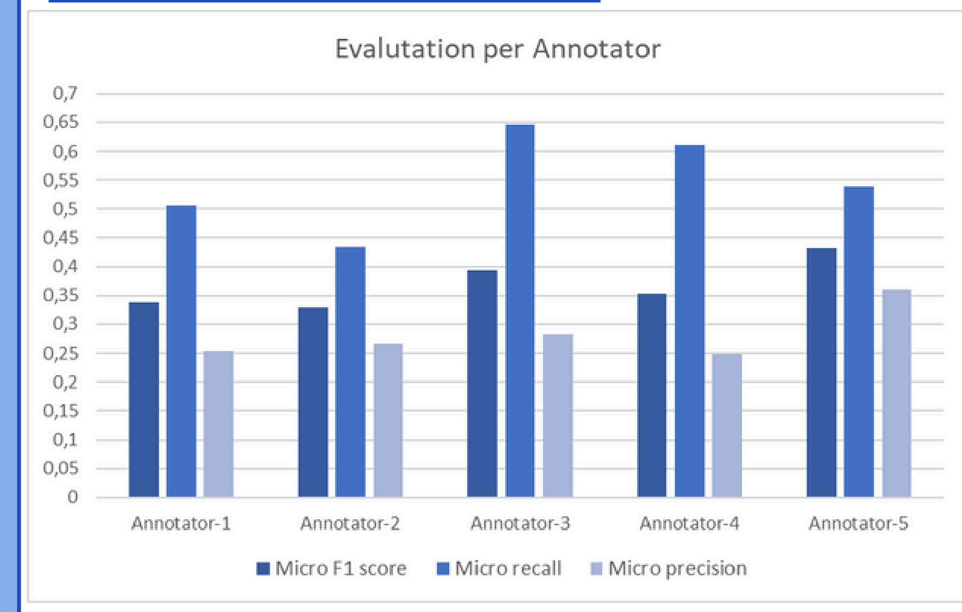


Graph 2: Oneshot soft label



Graph 3: Fewshot soft label

Hard Per Annotator Labels



Graph 4: F1 score, recall and precision per annotator

- **Annotated labels**: per annotator
- **Predicted labels:** >=2 LLM runs predicted it

5. Conclusion

Limitations:

- **Little** annotated **data** and **translated** from Dutch

To model an LLM to detect emotions

- Choose **existing model** (Llama3)
- Choose **prompting strategy** (zeroshot, oneshot, fewshot, CoT)

The effect of different prompting strategies

- **Oneshot** performed best in **recall**
- **Fewshot** performed best in **precision**
- **CoT zeroshot** had the **largest improvement**
- **Fewshot** is **not** enough to capture **annotator perspective**

The effect of different labels

- **Majority** hard labels allow **general predictions**
- **Per annotator** hard labels show **subjectivity** per annotator
- **Soft** labels allow for better more **precise subjective examples**

LLMs can predict emotions, as much "right" as the average annotator

6. Future Work

- Run code on **GoEmotions** dataset
- **Finetune** the model
- Try different **models** (e.g. Mistral, Zephyr, etc.)
- Find appropriate **evaluation measure** for **subjective** tasks

7. References

[1] E. Schneiderhan and K. Schamus, "Reasons and inclusion: The foundation of delib- eration", Sociological Theory, vol. 26, no. 1, pp. 1–24, 2008. doi: 10.1111/j.1467-9558.2008.00316.x.
[2] J. Forester, "Challenges of deliberation and participation", Les ateliers de l'éthique, vol. 1, no. 2, pp. 19–25, 2018. doi: https://doi.org/10.7202/1044678ar.
[3] E. Kelly and N. Kaminskien˙e, "Importance of emotional intelligence in negotiation and mediation", International Comparative Jurisprudence, vol. 2, no. 1, pp. 55–60, 2016. doi: 10.1016/j.icj.2016.07.001.
[4] K. Kim, N. Cundiff, and S. Choi, "The influence of emotional intelligence on negoti- ation outcomes and the mediating effect of rapport: A structural equation modeling approach", Negotiation Journal, vol. 30, no. 1, pp. 49–68, 2014. doi: 10.1111/nejo. 12045
[5] A. S. Cowen and D. Keltner, "What the face displays: Mapping 28 emotions conveyed by naturalistic expression", The American psychologist, vol. 75, no. 3, pp. 349–364, 2020. doi: 10.1037/amp0000488.
[6] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and R. Sujith, "Goemotions: A dataset of fine-grained emotions", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4040–4054, 2020. doi: 10.18653/v1/2020.acl-main.372.
[7] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, "Learning from disagreement: A survey", Journal of Artificial Intelligence Research, vol. 72, pp. 1385–1470, 2021. doi: https://doi.org/10.1613/jair.1.12752

TU Delft Contact info: b.c.p.zuurbier@student.tudelft.nl