Leveraging Large Language Models for Classifying Deliberative Elements in Public Discourse

Author: Adina Dobrinoiu, a.n.dobrinoiu-1@tudelft.nl

Responsible professor: Luciano Cavalcante Siebert
Supervisors: Amir Homayounirad, Enrico Liscio

1. Introduction

- **Public deliberation** is a way in which citizens can exchange opinions and discuss problems in detail in a respectful and reasoned manner [1].
- The **attainability and effectiveness of deliberation**, both in theory and practice, is **based on argument formalization** [2].
- However, subjectivity is an inherent challenge in deliberation.
- Other difficulties associated with deliberation are the **large volumes of data** produced in such debates [1] and the **low accuracy of results**, partly attributed to low participation rates [1].

2. Research question

Can LLMs detect the subjective arguments that support different stances in a deliberation?

Subquestions:

- How can LLMs flag and classify subjective arguments in public discourse?
- What evaluation metrics can be used to assess the performance of LLMs in argument extraction?
- How do few-shot and zero-shot approaches compare, and what impact does adding chain-of-thought reasoning have on their performance?

6. References

[1] R. Shortall, A. Itten, M. v. d. Meer, P. Murukannaiah, and C. Jonker. Reason against the machine? future directions for mass online deliberation. Frontiers in Political Science, 4, October 2022.
[2] J. Fishkin and R. Luskin. Experimenting with a democratic ideal: Deliberative polling and public opinion. Acta Politica, 40:284–298, 2005.
[3] T. Govier. ”A practical study of argument”, Cengage Learning, 2013, 1-21

3. Methodology

- **Generate labels for the dataset**
  - An argument is a set of claims where **premises** provide reasons for the **conclusion** [3].
  - What this research is then concerned with is **extracting the premises of an argument.**

zero-shot

one-shot **Llama 2** **few-shot**

+

chain-of-thought reasoning

4. Results

Subjectivity in annotation

Pairwise F1 score

average: 0.2447
max: 0.4
min: 0.11

Annotators	1	2	3	4	5
1	1	0.55	0.71	0.71	0.69
2	0.55	1	0.6	0.59	0.6
3	0.71	0.6	1	0.68	0.68
4	0.71	0.59	0.68	1	0.69
5	0.69	0.6	0.68	0.69	1

Figure 3: The one-shot pairwise cosine similarity score between the different LLM responses for each annotator

Annotators	LLM method - removed data improvement		
	Zero-shot	One-shot	Few-shot
Annotator 1	0.297	0.299 (+0.056)	0.270 (+0.009)
Annotator 2	0.086	0.051 (+0.000)	0.104 (+0.000)
Annotator 3	0.283	0.321 (+0.082)	0.283 (+0.024)
Annotator 4	0.233	0.260 (+0.046)	0.217 (+00.09)
Annotator 5	0.161	0.173 (+0.026)	0.159 (+0.002)

Figure 1: Cosine similarity evaluation on LLM prompting approaches after overfitted data is removed

Annotators	LLM method – chain-of-thought improvement		
	Zero-shot	One-shot	Few-shot
Annotator 1	0.233 (-0.064)	0.225 (-0.074)	0.245 (-0.025)
Annotator 2	0.374 (+0.288)	0.147 (+0.096)	0.174 (+0.070)
Annotator 3	0.318 (+0.035)	0.361 (+0.040)	0.311 (+0.028)
Annotator 4	0.234 (+0.001)	0.222 (-0.038)	0.207 (-0.010)
Annotator 5	0.271 (+0.110)	0.168 (-0.050)	0.143 (-0.016)

Figure 2: Cosine similarity evaluation on LLM prompting approaches after chain-of-thought reasoning is applied

5. Conclusions

- The data annotation task proved to be subjective by the low interrater agreement score and the variance in how many data entries each annotator considered not to be arguments.
- In one-shot and few-shot approaches, the LLM overfit the examples in the prompts, leading to unexpectedly better performance in zero-shot.
- Chain-of-thought reasoning proved to be efficient for argument identification
- Pairwise cosine similarity scores showed LLM responses aligned more with annotators sharing similar labels.

Future work:

- Implementing automatic key point extraction to generate a set of labels to be used for annotation.
- Generating more annotations for the dataset.