Leveraging LLMs for Classifying Subjective Topics Behind Public Discourse

Author: Ana Cristiana Marcu <a.c.marcu-1@student.tudelft.nl>
Supervisors: Amir Homayounirad, Enrico Liscio
Responsible Professor: Luciano Cavalcante Siebert

TU Delft

1. Introduction

Public deliberations
- a vital component of the democratic system [1]

Challenge
- **unstructured nature of deliberations** challenges moderators to comprehend and analyze the large volume of data produced [2]

First step in structuring deliberations
- identifying **topics** -> *multi-label classification problem*

Further challenges
- **labeled data** necessitates employing a group of annotators -> process that is both *costly* and *time-consuming*
- **annotator's disagreement** [3]

Possible Solution
- **LLMs** offer a promising opportunity to revolutionize the *identification of subjective data annotation*

2 core objectives:
1. Identifying Gold Label
2. Exploring Subjective Human Labels

2. Research questions

How can Large Language Models classify subjective topics behind public discourse?

3. Data

Dataset: Energy in Súdwest-Fryslân case study
482 responses

Label extraction: BERTopic [4] -> 6 labels

Data annotation: 5 annotators
50 data Items
- overall moderate agreement (based on Fleiss Kappa metric)

Data aggregation: majority vote (>50%)
no aggregation

4. Methodology

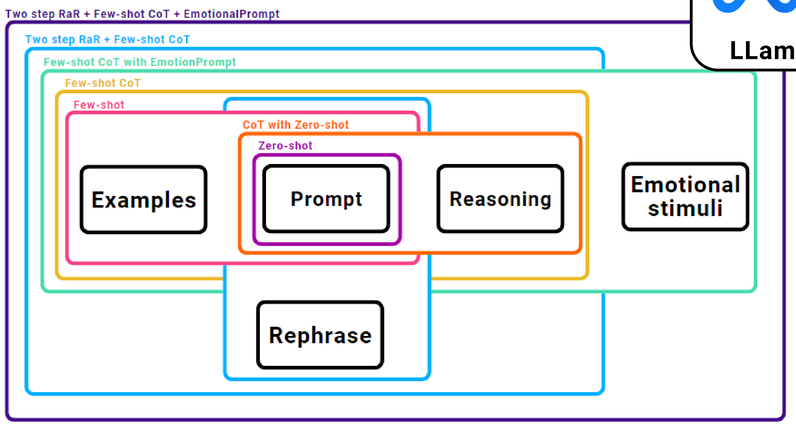

Figure 1: Overview of Prompting Strategies

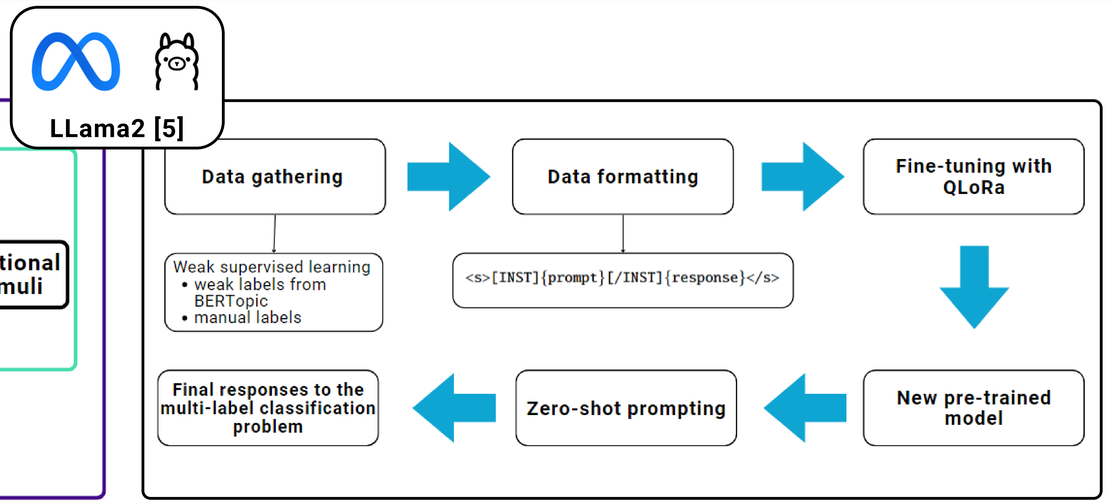

Figure 2: Process of extracting topics - from data gathering to Fine-tuning LLaMa-2 with QLoRa [6]

5. Results and Discussion

- **Each method -> run 10 times and aggregates using MV**
- **Preprocessing results using Sentence Transformers**

Prompting/training method	Micro F1-Score
Zero-shot	0.64
Zero-shot CoT	0.657
Few-shot	0.817
Few-shot CoT	0.817
Fine-tuning with QLoRa	0.865

Table 1: Micro-F1 Score Results for Identifying Gold Labels

Prompting method	Averaged Micro F1-Score for all annotators
Few-shot	0.756
Few-shot CoT	0.715
Few-shot CoT v2	0.75
RaR + Few-shot CoT v2	0.682
Few-shot CoT v2 + EmotionPrompt	0.782
RaR + Few-shot CoT v2	0.779

Table 2: Averaged Micro-F1 Score Results for Prompting Methods for Exploring Subjective Human Labels

Limitations
- **hallucination [7] (especially for CoT method)**
- **dependency on high-quality data (fine-tuning and evaluation)**
- **low number of annotations and not a diverse pool of annotators**

6. Conclusion and Future work

1. The potential of LLMs to identify subjective topics behind public discourse has been highlighted through the study

2. Identifying Gold Label
- Fine-tuning LLama-2 with QLoRa (best Micro-F1 score)

3. Exploring Subjective Human Labels:
- Few-shot CoT v2 + EmotionPrompt [8] (best Micro-F1 score)

Future work
1. Expand the annotated dataset
2. Expand the pool of annotators to be more diverse
3. Fine-tune LLM for Exploring Subjective Human Labels
4. Explore the hallucination issue
5. Different temperature settings
6. Soft probabilistic labels
7. Explore the use different LLMs

References

[1] J. S. Fishkin, Democracy and deliberation: New directions for democratic reform. Yale University Press, 1991.
[2] 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, vol. 4, p. 946589, 2022.
[3] N. Deng, S. Liu, X. F. Zhang, W. Wu, L. Wang, and R. Mihalcea, "You are what you annotate: Towards better models through annotator representations," arXiv preprint arXiv:2305.14663, 2023.
[4] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," arXiv preprint arXiv:2203.05794, 2022.
[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.
[6] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoye "Qlora: Efficient finetuning of quantized llms,"Advances in Neural Information Processing Systems, vol. 36, 2024.
[7] H. Duan, Y. Yang, and K. Y. Tam, "Do llms know about hallucination? an empirical investigation of llm's hidden states," arXiv preprint arXiv:2402.09733, 2024.
[8] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie, "Large language models understand and can be enhanced by emotional stimuli," arXiv preprint arXiv:2307.11760, 2023.