Decoding Sentiment with Large Language Models

Timur Oberhuber,[@,*] Supervisors: L. C. Siebert,[*] A. Homayounirad,[*] E. Liscio[*]

[@]t.oberhuber@student.tudelft.nl, [*]EEMCS, Delft University of Technology

Introduction & Background

Context - Public Deliberation: A value-based discussion that includes ordinary people, especially marginalized groups, to find transformative solutions to social problems [1].

Problem - Subjective Sentiment: Effective deliberation needs one moderator per twenty participants [2], each with their own "subjective sentiment" (personal feelings, views, and beliefs) [3]. The need for many moderators hinders scaling public deliberation.

Solution? - Sentiment Analysis: Extracts opinion polarity from text (positive, negative, or neutral). This can give moderators a better overview of participants, allowing for more participants per moderator.

Research Question

Can a Large Language Model (LLM) detect subjective sentiment of statements within the context of public deliberation?

Methodology

LLM: Llama 3 running on Python 3.12.3 in Ollama 0.1.8.
Data: Textual opinions of 1376 Sudwest-Frysland residents on future energy policy [4], annotated with sentiment by 5 peers.
Prompting Strategies:

- **Zero-Shot:** Directly predicts sentiment without any task-specific examples.
- **Few-Shot:** Provided with a few examples of text-sentiment pairs prior to making predictions.
- **Chain-of-Thought (Zero- & Few-Shot):** Guided to reason step-by-step about its predictions. Zero-shot uses no examples; few-shot modifies training data for reasoning.

Scenarios:

- **Hard Label:** Assigns a single label based on the majority of annotators.
- **Soft Label:** Averages labels from multiple annotators to provide a confidence level instead of a single label [5].
- **Subjective Label:** Captures each annotator's perspective when predicting labels.

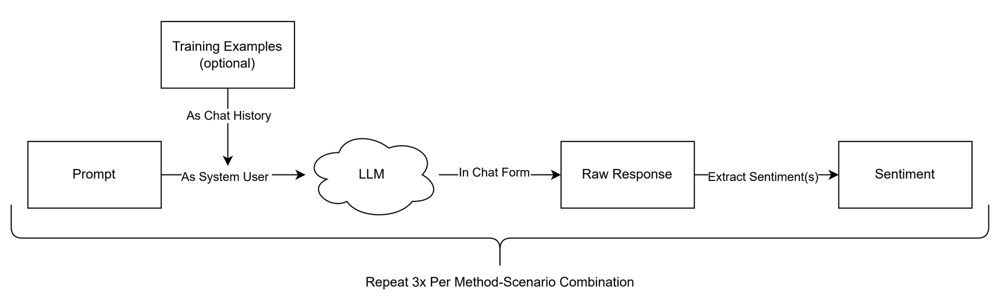

Figure 1: An Overview of the Methodology.

Results

Fleiss' Kappa (10,000 Splits): 0.17 (*p-value* of 1.0)
One-Way ANOVA: *p-value* of 0.639

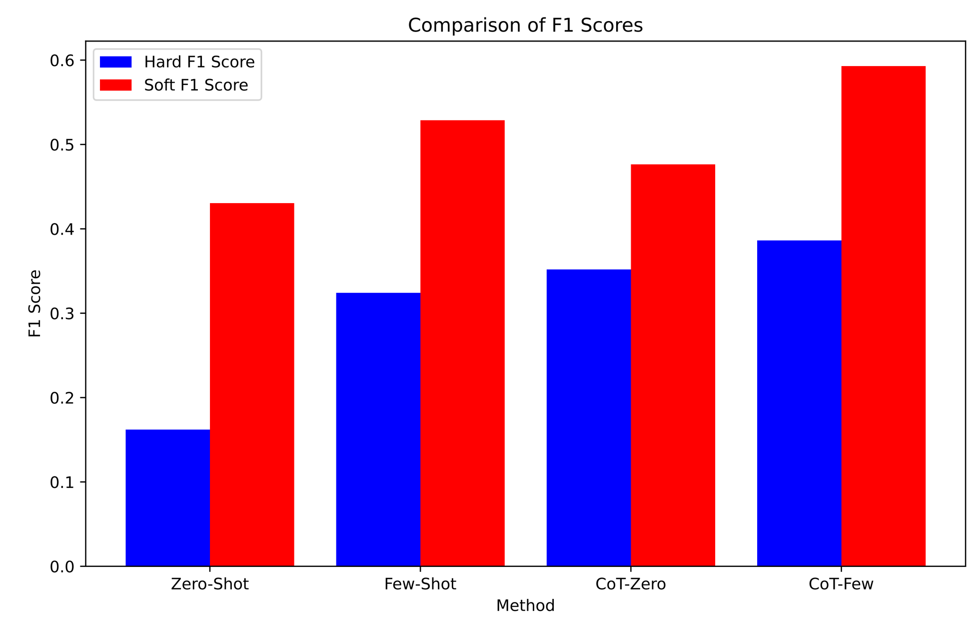

Figure 2: F1-Scores Across Soft/Hard Label Scenarios for All Methods.

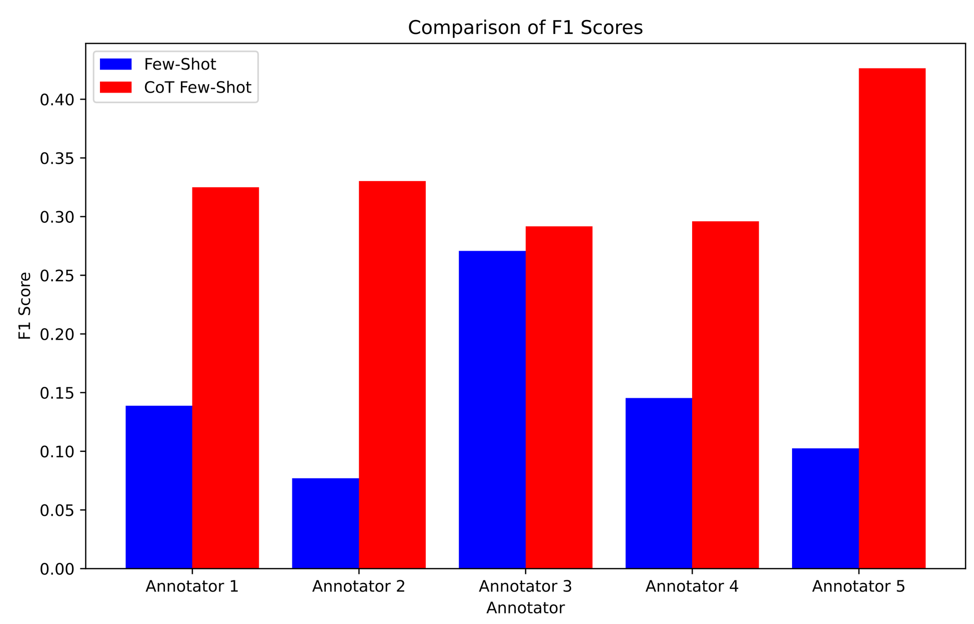

Figure 3: F1-Scores by Annotator and Method in the Subjective Scenario.

Accuracy: $0.1 <$ F1-Scores for All Methods, in All Scenarios.

Conclusions

- LLMs can detect subjective sentiment in public deliberation.
- LLMs shouldn't replace human judgment.
- Sentiment isn't always binary, as shown by the metrics of the soft-label scenario.
- Combining reasoning capabilities and training data (CoT Few-Shot) is most effective.

Future Work

- Replicate the study with larger datasets from different contexts.
- Improve annotator agreement with objective guidelines.
- Investigate reinforcement learning with human feedback.
- Conduct a longitudinal study on integrating sentiment analysis in public deliberation tools.

References

[1] E. Blacksher, A. Diebel, P.-G. Forest, S. D. Goold, and J. Abelson, "What is public deliberation," *Hastings Cent Rep*, vol. 42, no. 2, pp. 14–17, 2012. [Online]. Available: https://doi.org/10.1002/hast.26

[2] M. Klein, "Enabling large-scale deliberation using attention-mediation metrics," *Computer Supported Cooperative Work (CSCW)*, vol. 21, pp. 449–473, 2012. [Online]. Available: https://doi.org/10.1007/s10606-012-9156-4

[3] B. Liu, *Sentiment analysis and opinion mining*. Springer Nature, 2022. [Online]. Available: https://www.cs.uic.edu/ liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[4] S. L. Spruit and N. Mouter, "Energy in súdwest-fryslân," 2020. [Online]. Available: https://www.tudelft.nl/en/tpm/pve/case-studies/energy-in-sudwest-fryslan

[5] N. Vyas, S. Saxena, and T. Voice, "Learning soft labels via meta learning," 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2009.09496

TUDelft