Leveraging LLMs for subjective value detection in arguments

How can Large Language Models (LLMs) be utilized to detect the subjective values of subjective arguments in public discourse?

Authors
Joosje Gorter
Email: J.C.E.Gorter@student.tudelft.nl

Supervisors
Luciano Cavalcante Siebert
Amir Homayounirad
Enrico Liscio

01 Introduction

Public deliberation is filled with argumentative statements, and identifying the underlying values driving these statements can be key to constructive discussions. Human values, which are beliefs guiding behavior and decision-making, play a crucial role in public discourse. They are essential in public discourse and deliberation because they provide the ethical framework that guides decision-making. They characterize societies and individuals, and a balance between these values is crucial for fostering harmonious and constructive public deliberation. Thus, a better understanding of values can facilitate more productive discussions and potentially lead to better outcomes.
This research aims to explore the different ways of utilizing LLMs to detect these underlying values.

02 Research Questions

"How can LLM's be utilized to detect the subjective values of arguments in public discourse?"

Sub-questions:
- How can the underlying values of subjective statements be annotated?
- What kinds of methods are used for LLM utilization?
- How can performance of value detection be measured?

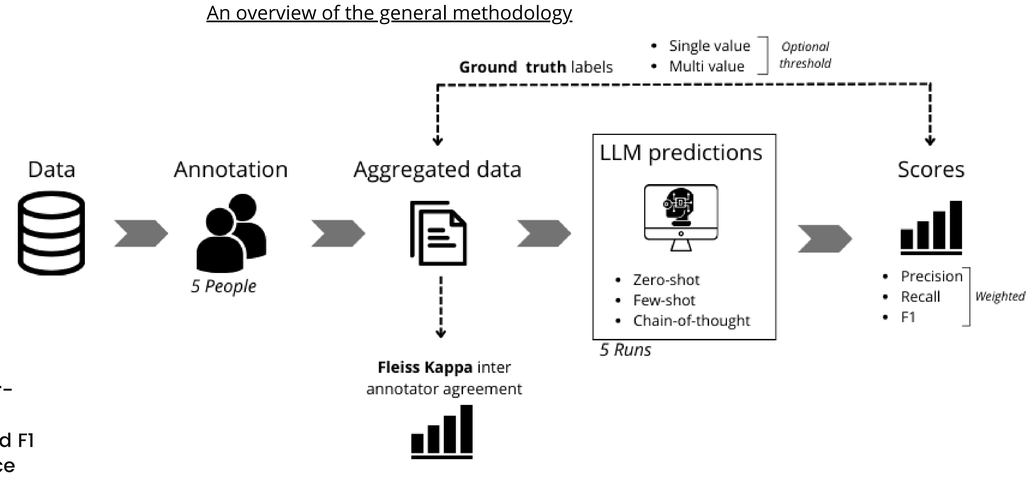
03 Methodology

Data
The dataset used is from a case study on public deliberation in Súdwest-Fryslân [1]. The data was annotated by 5 students.

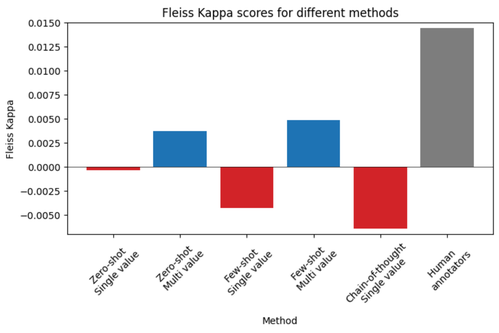
Prompting methods
- Zero-shot
- Few-shot
- Chain-of-thought

Metrics
- Fleiss Kappa statistic for inter-annotator agreement
- Weighted precision, recall and F1 scores for overall performance

An overview of the general methodology



04 Results



Low Fleiss Kappa scores indicate low consensus over multiple runs of the LLMs, indicating inconsistent replies

Performance observations
- Single-value prediction using chain-of-thought prompting is the best-performing method, with an F1 score of **0.594**
- There is a large difference in performance between single- and multi-value predictions.
- Overall, with F1 scores not surpassing 0.6, the performance is suboptimal. However, this could be due to the metrics used not being suitable for the task. Research by Uma et al.[3] suggests that a multi-label approach combined with a soft-loss function could yield better results.

Method	LLM prediction	Weighted Precision	Weighted Recall	Weighted F1
Zero-shot	Single value	0.596	0.597	**0.567**
	Single value (Threshold = 3)	0.555	0.544	0.525
	Multi value	0.382	0.167	0.199
	Multi value (Threshold = 3)	0.233	0.029	0.029
Few-shot	Single value	0.621	0.610	**0.587**
	Single value (Threshold = 3)	0.607	0.590	0.570
	Multi value	0.349	0.169	0.193
	Multi value (Threshold = 3)	0.007	0.029	0.011
Chain-of-thought	Single value	0.620	0.603	**0.594**
	Single value (Threshold = 3)	0.580	0.603	0.594

Overall scores of different prompting methods when compared to human annotations

05 Conclusion

1) When only considering F1 scores, one could conclude that LLMs are not yet equipped to accurately predict subjective underlying human values in argument statements. However, as mentioned in the results section, a soft-loss function approach for the multi-value predictions might yield different results.

2) There is a lack of suitable metrics for evaluating the performance of LLMs on highly subjective tasks such as value detection. Traditional hard-metrics, such as the F1 score, show poor performance. However, research suggests that when multi-value ground truth labels are combined with a soft-loss function, they should outperform single majority labels [3]. Given the potential increase in measured performance for multi-label tasks and the unsuitability of metrics such as ranking comparisons, further research on metrics for evaluating highly subjective tasks is strongly recommended.

Related Literature

[1] "Energy in Súdwest-Fryslân," TU Delft. Accessed: May 15, 2024. [Online]. Available: https://www.tudelft.nl/en/tpm/pve/case-studies/energy-in-sudwest-fryslan

[2] J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth, and B. Stein, "Identifying the Human Values behind Arguments," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4459–4471. doi: 10.18653/v1/2022.acl-long.306.

[3] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, "A Case for Soft Loss Functions," Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 8, pp. 173–177, Oct. 2020, doi: 10.1609/hcomp.v8i1.7478.