Leveraging LLMs for subjective value detection in argument statements

Joosje Gorter

Supervisors: Luciano Cavalcante Siebert, Amir Homayounirad, Enrico Liscio

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Joosje Gorter
Final project course: CSE3000 Research Project
Thesis committee: Luciano Cavalcante Siebert, Amir Homayounirad, Enrico Liscio, Jie Yang

Abstract

This paper investigates the use of Large Language Models (LLMs) for automatic detection of subjective values in argument statements in public discourse. Understanding the underlying values of argument statements could enhance public discussions and potentially lead to better outcomes. The LLM utilization methods tested were zero- and few-shot prompting, as well as chain-of-thought prompts. In order to compare the predictions made by the LLM, a set of ground truth labels was required as an established baseline. For these labels, either single majority labels or multi-value labels were considered, both derived from a set of aggregated human annotations. Results indicated that LLM performance was sub optimal, achieving a maximum weighed F1 score of 0.594 for single-value chain-of-thought predictions. Additionally, current metrics were found inadequate for assessing LLM performance on a highly subjective task such as value detection, evidenced by poor scores in multi-value predictions despite subjective evaluation suggesting otherwise. Furthermore, a last experiment was aimed at capturing a specific annotator's subjectivity. This yielded inconsistent results, with f1 scores peaking around 0.4, indicating that LLMs are not well-suited for emulating individual human subjectivity.

1 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. According to Bleksher et al.[1], citizens must "identify, clarify, and weigh the tensions among their views and the values underlying them; justify them to others; and set priorities". Human values are essential in public discourse and deliberation because they provide the ethical framework that guides decision-making. They characterize societies and individuals [2], 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. Recent years have seen a spike in the usage of Large Language Models (LLMs). While they have become integral to our everyday interactions, their performance on tasks that even humans might disagree on remains a question [3].

The main research question that will be considered in this paper is '*How can LLM's be utilized to detect the subjective values of arguments in public discourse?*'. This is then broken down into the following 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?

This research aims to explore the different ways of utilizing LLMs to detect these underlying values in argumentative statements. Building upon the work of Kiesel et al.[4], who developed a taxonomy of 54 value labels and used them for a multi-label classification problem, this paper seeks to contribute to the research on value detection in argument statements by further investigating different methods of LLM utilization. These methods consist of zero-shot prompting, few-shot prompting, and chain-of-thought reasoning prompts. Additionally, this research investigates how defining ground truth labels for human annotation affects the performance of LLMs in automatic value detection.

2 Background

Tthere are two key concepts that should be explained: *Public discourse* and *the definition of a value*. This section provides these definitions and highlights the role that values play in public discourse. Lastly, it mentions the related work upon which this paper builds.

2.1 Public discourse

Public discourse refers to public discussions that seek collective solutions to challenging social problems [1].
Seeking these collective solutions through free and open discussion of public policies is what our democratic society is based on [5], as they are an important part of policy making. Citizen engagement in these debates is critical, as they might see policies as more acceptable and fair when they are subject to an open, inclusive discussion, even if they might disagree with the decision [6].

2.2 Definition of a value

The difficulty of defining what human values are lies in the extremely high level of subjectivity. "There are no objective values." As philosopher John Leslie Mackie states in his essay on the subjectivity of values [7]. This research uses the work of Schwartz et al., who have striven find a definition for human values that individuals in all cultures recognize [2].

Schwartz defines a value as "a (1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities."[8]

Consider the following example:

> *"We should not spend money on installing solar panels. It might be good for the environment, but the financial trade-off is not worth it."*

In this case, the desirable outcome (2) is to not spend too much money, as they generally deem the financial trade-off not worth the money (3). Thus, the choice is to not invest in solar panels (4), as they prefer the value of *Have wealth* over that of *Be protecting the environment* (5).

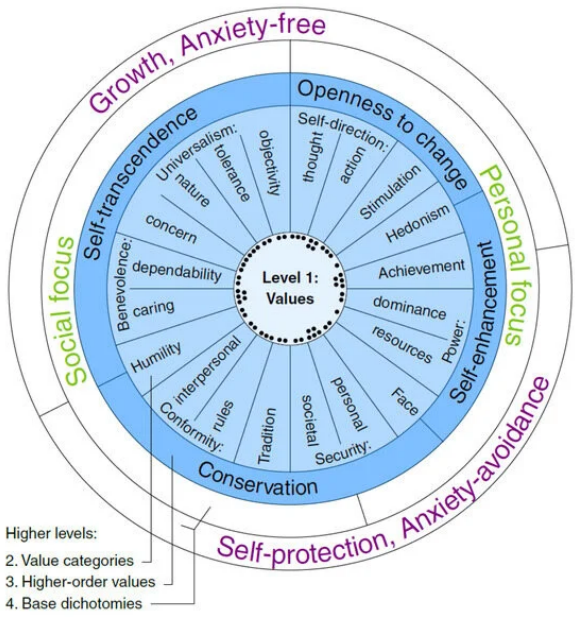

Figure 1: Taxonomy of values. From *Identifying the Human Values behind Arguments* by Kiesel et al. (2022)

2.3 Related work

The approach of using LLMs for subjective annotation tasks within the Social Sciences has been explored by Weber and Reichardt. The results highlighted "the need for careful validation and tailored prompt engineering"[9]. While their research also explores the utilization of LLMs for subjective annotation tasks, this paper aims to expand upon the various prompting strategies that can be used. While Weber and Reichardt's research examines the utilization of LLMs for a broad range of annotation tasks within the social sciences, this paper focuses on various prompting strategies that can be employed. In contrast to their generalized approach, this research focuses specifically on the task of detecting subjective human values.

Previous work by Liscio et al. has addressed this task, with one paper discussing the cross-domain classification of moral values in text using the language model BERT [10], and another exploring a hybrid (human and AI) approach to identify context-specific values [11]. The paper on cross-domain classification explores the generalizability of BERT's capabilities in value detection. Their research was done by pre-training the model, whereas this paper focuses solely on the effect of different prompting strategies and the difference between single- and multi-value prediction.

Given the similar nature of the research, the main reference for this study is the paper *Identifying the Human Values behind Arguments* by Johannes Kiesel et al [4]. They were the first to try detecting human values in argument statements using LLMs. For this, they used a taxonomy of 54 human values as composed in the earlier work of Schwartz et al [12]. An overview of this taxonomy, which will also be used in this research, can be seen in figure 1.

3 Data

The dataset used in this research was taken from a consultation of 1376 residents of the municipality of Súdwest-Fryslân (South West Friesland) on the future energy policy of their municipality [13]. Since it is not a public dataset, all LLM interaction was done locally.

The dataset consists of argument statements supporting residents' rankings of six possibilities for future municipal energy policy. Therefore, this dataset is well-suited for detecting values in public discourse. A potential drawback of the dataset is that the original language is Frysian, and the available Dutch and English translations were generated automatically, potentially introducing translation errors [14].

3.1 Annotation

The dataset was manually annotated by 5 computer science students. This was done by providing them with the list of 54 values from Kiesel et al [4] and having them select all values they think apply to an argument statements. In total, 50 data points were annotated. These results were then aggregated such that every argument statement had a list of labeled values and their total annotation frequency. This aggregation was necessary to allow for a multi-run analysis of LLM performance, thereby reducing the issue of observed variance between individual runs.

3.2 Annotator agreement

To determine the inter-annotator agreement, the Fleiss Kappa statistic was used. This statistic is suitable for situations where there are more than two annotators, as it takes into account the amount of agreement that is expected between annotators purely by chance [15].
Fleiss' Kappa is defined as

$$k = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{1}$$

Where the denominator refers to the degree of agreement that is possible above chance, and the numerator is the actually achieved agreement above chance. A k of 1 indicates complete agreement among annotators, and a k below 1 indicates that there is no agreement.

The Fleiss Kappa score for the aggregated human annotations was **0.0144**. Following the interpretation provided by Landis and Koch [16] this indicates slight agreement among annotators. A slight note should be added here, as there is no universally accepted interpretation of the Kappa Fleiss statistic. However, most interpretations found in literature would suggest that this score can be seen as having only slight agreement among annotators. [17]

4 Methodology

This section outlines the methods used in this research. It mentions the prompting methods used for LLM utilization, the chosen evaluation metrics, the compositions of ground truth labels and lastly it explains the extra experiment to capture person-specific subjectivity. The main methodology can be seen in figure 2.

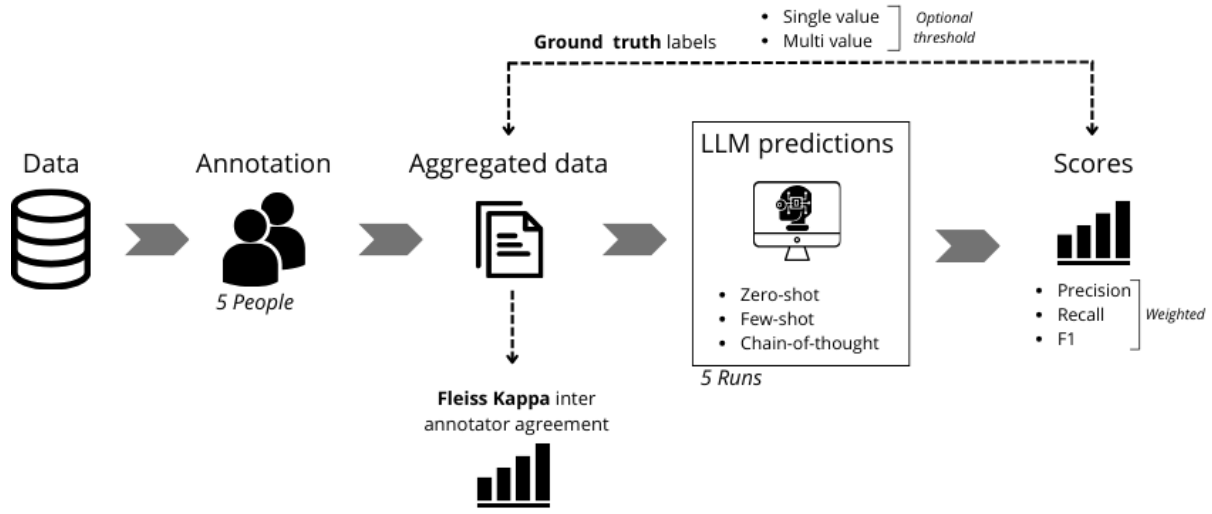

Figure 2: Overview of general methodology

4.1 LLMs and their utilization methods

Due to the dataset not being public, as described in chapter 3, the 8B parameter version of LLama3 was used locally using Ollama [18]. The three methods of LLM prompting that were used were

- Zero-shot prompting
- Few-shot prompting
- Chain-of-Thought prompting

For both methods the LLM was asked to pick from the list of 54 provided values to select the value found most fitting for the given argument statement. The difference between zero-shot and few-shot prompting is that in the former no examples are provided, whereas in the latter 3 examples of argument statements and its annotated value are given in the prompt [19].

For chain-of-thought prompting, the model is first asked to provide its reasoning for performing a certain task, and from that reasoning it is asked to predict a value. In this case, the first prompt asks the model to reason about which values it finds suitable for a given argument statement. The second prompt then feeds the given reasoning back into the model, and asks it to predict a value, given that reasoning. Chain-of-thought prompting has shown to increase performance on a wide range of tasks [20].

4.2 Evaluation metrics

Multiple metrics were utilized to evaluate the model's performance in automatic value prediction, as detailed in table 1. Given the large number of labels, weighted averages were analyzed. The micro-average was considered sub-optimal for this task because it disregards true negative predictions. Accurately excluding the majority of labels (true negatives) is indicative of accurate performance.

4.3 Ground truth labels

In order to compare the predictions made by the LLM, a set of ground truth labels was required as an established baseline. Due to the subjective nature of the task of detecting underlying human values of argument statements, a big challenge in this research was to determine what kind of label to use as the ground truth. Three methods were tried; Majority labels, Majority labels with threshold and Soft labels.

Majority label

To simplify the process of evaluating the model performance, the first ground truth labels used were single majority labels. The reasoning for this was that single value predictions were expected to be the easiest for LLMs to execute. Another consideration was that due to the lower complexity of the task, the execution times would be fast. These majority labels were determined by taking a majority vote among the 5 annotators and taking the value with the most annotations as the single ground truth label for that argument statement. On the occurrences where multiple values would be contenders for the majority label, one was chosen at random.

Majority label with threshold

Since there was lacking consensus among the annotators, and a high disparity of annotated values (on average 10,78 unique values were assigned to each argument statement), a threshold seemed necessary to ensure the ground truth labels reached at least a baseline of consensus. For this the aggregated annotations were re-evaluated, and only if at least 3 out of 5 annotators had picked a value was it a viable option as a majority label. This meant that some ground truth labels did not meet this threshold and were assigned 'None'. To accommodate this option, the LLM prompt was changed to allow 'None' if the model found no value applicable. An important note here is that the LLM never actually picked 'None' as an option, meaning that the expected increase in performance did not actually occur.

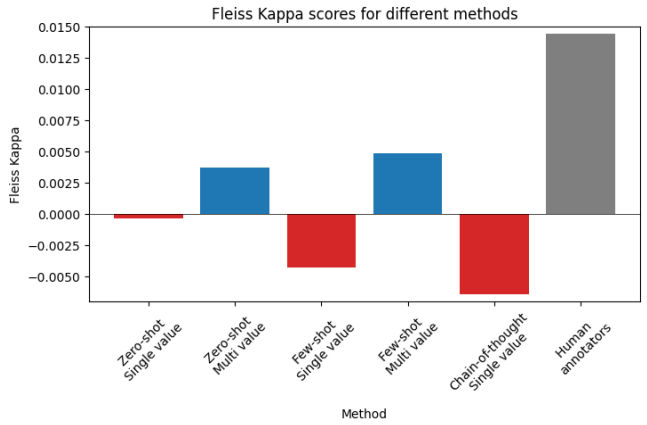

Figure 3: Fleiss Kappa scores over 5-run aggregated predictions for different prompting methods (single and multi value predictions)

Soft labels

The single majority label seemed unsuitable for this highly subjective task. Seeing as us humans had a hard time choosing a single value, it was a hard ask to expect the LLM to do so. To accommodate a more accurate representation of the human task, the LLM was asked to select any values it inferred to be applicable. This prompt was then ran 5 times over all data points, and the results were aggregated to mimic the 5 human annotators.

4.4 Capturing subjectivity

A last experiment that was conducted was to do annotator-specific predictions. The aim of this was to determine if an LLM is able to capture a specific person's subjectivity. This was done by providing the LLM with 30 examples of an annotator's annotations, after which the remaining 20 data points were used to prompt the LLM to predict which values that annotator would choose. For consistency, the same metrics were used to show the performance of the LLM on this task.

5 Results

This section presents the findings of this study. It first provides an overview of the Fleiss Kappa scores for all prompting methods, indicating the inter-annotator agreement over multiple runs. Secondly it shows the general scores achieved by multiple methods, using the metrics introduced in section 4.2, after which the effect of using a threshold for ground truth labels is mentioned. The fourth subsection briefly mentions the difference in scores between zero- and few-shot prompting, which will be further addressed in the discussion in chapter 6. Lastly, the results of the experiment on capturing annotator-specific subjectivity are presented.

5.1 Fleiss Kappa scores

An overview of the Fleiss' Kappa scores for all prompting methods, for both single- and multi-value LLM predictions, is presented in Figure 3. These scores were calculated on the aggregated predictions of running the model 5 times. The results for single-value LLM predictions show negative values for zero-, few-shot and chain-of-thought prompts, indicating poor or no agreement. In contrast, the multi-value predictions display positive values, suggesting slight agreement among annotators. This indicates, that when given the option to select multiple values, the LLM is more consistent in its selection of values compared to when having to select a single value.

5.2 General scores

Table 1 displays the weighted precision, recall and F1 scores for all prompting methods. Most noticeable here is the large difference in scores between single- and multi-value predictions, with the former displaying the highest scores. The difference in performance between prompting for single-value predictions is slight, with zero-shot having a weighted F1 score of 0.567 and chain-of-thought scoring the best out of all methods with 0.594.

5.3 Effect of threshold

The application of an annotation and prediction threshold to determine the eligibility of values as ground truth labels also had an impact. With the use of a threshold of 3 slightly impacting the F1 performance in a negative manner, with threshold scores being consistently lower across all methods.

This is likely due to the lower degree of agreement between LLM responses, when compared to the human annotators, as shown in Figure 3. It is possible that this caused a discrepancy in the number of values included in the ground truth labels, which could account for the slight drop in performance.

5.4 Zero- vs Few-shot

It is also interesting to note that the difference between the best scores of zero-shot and few-shot is not as large as one might expect. With zero-shot prompting achieving a maximum weighted F1 score of 0.567 and few-shot peaking at 0.587. As of yet, the reason for this discrepancy remains unclear, although it might be due to the selected examples provided in the prompt. This indicates that the addition of only a few samples (3 were used in this experiment) does not provide any performance gain on such a subjective task.

5.5 Per-annotator predictions

The results of tasking the LLM with annotator-specific predictions can be found in figure 4. An important thing to note here is that the results differ greatly over runs. For example, this is a list of weighted F1 scores per run for a 5 run aggregation of annotator 2: [0.159, 0.111, 0.173, 0.188, 0.187]. One run scored as low as 0.111, while the average score over the 5 runs was 0.164. This shows great inconsistency in how well the LLM is able to perform this task.

6 Discussion

This section discusses some of the observed results and explores possible implications. Firstly, it touches upon the difference in performance between single- and multi-value predictions. Secondly, it notes on the overall scores achieved and provides an assessment on the suitability of LLMs for predicting subjective human values in argument statements.

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

Table 1: Scores for LLM predictions when compared to human annotation

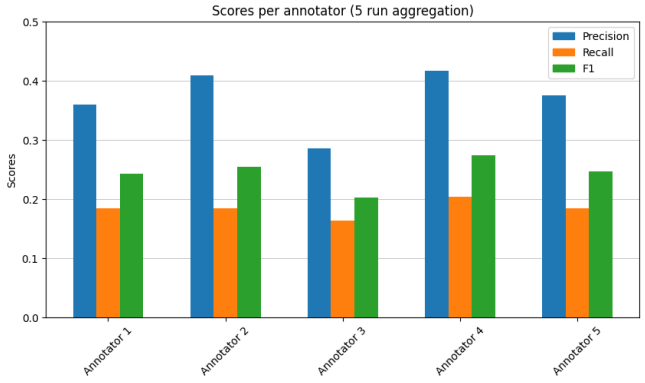

Figure 4: Scores for annotator-specific multi-value predictions

6.1 Performance of single- vs multi-value prediction

As mentioned in section 5.2, there is a large difference in performance between single- and multi-value predictions, with multi-value predictions scoring significantly lower. This large difference is surprising, seeing as works by Uma et al. [3], [21] show a better performance when using a multi-label ground truth when compared to using majority voting. The low scores for multi-value predictions are liekly due to the loss function used in this study. Both single- and multi-value predictions were scored using the same metrics (precision, recall and F1) to allow for a direct comparison between the two approaches, whereas the work by Uma et al. [21] found that a combination of probabilistic soft labels with a probability comparing soft-loss function worked best on a broad range of tasks.

Options for soft-loss functions were explored, but ranking comparisons, such as Kemeny-Young [22], require a definitive ordered ranking and an equal amount of ranked items. When assigning values to an argument statement, it is impossible to definitively determine whether a given value played a driving role due to the subjective nature of human values, as discussed in section 2.2. Consequently, it is difficult to ascertain the number of values that can be assigned to a statement, as the lack of objectivity in individual cases precludes a clear numeric boundary.

Even when using hard metrics for comparison between single- and multi-value predictions, the large difference was unexpected, as the LLM's performance did not appear particularly poor throughout this study. In fact, distinguishing between the aggregated results of LLM responses and human annotations might be challenging for most people. The foremost noticeable difference is in the average number of unique values assigned per argument statement, with LLMs averaging 4.52 unique values and humans up to 10.74, given a team of five annotators. Despite this significant difference in annotation composition, subjective observation does not suggest either approach being superior, even though current metrics might indicate otherwise. This discrepancy between the observed quality of LLM predictions and their performance measured by hard metrics could be an area of future research, with the suggestion that subjective tasks might benefit from more subjective evaluation methods.

6.2 Assessment of LLM utilization for value detection

The results shown in section 5.2 indicate that chain-of-thought prompting used for single-value prediction is the best performing method. With a weighed F1 score of 0.594 it outperformed the zero- and few-shot prompting methods. Although the chain-of-thought prompting methods performed best out of the methods tested, it did not perform particularly well. None of the methods managed an F1 score of over 0.6, which indicates poor performance overall.

The low scores observed in this study can be attributed to two main factors. Firstly, the high number of unique values typically annotated by humans (10.74) indicate that detecting underlying human values in argument statements should not be treated as a single-label classification task. Due to its

highly subjective nature, consensus on a single value is rare. Secondly, the difficulty of comparing subjective multi-value predictions, as discussed in section 6.1, impacts the scores for multi-value predictions. A multi-label approach combined with some sort of soft-loss function would be optimal, as found by Uma et al. [21]. This suggests that the hard-metric comparison used in this study was not the most suitable for this task.

Based on the observed F1 scores alone, one might conclude that LLMs are not (yet) equipped to accurately predict subjective underlying human values in argument statements in public discourse. However, as previously discussed, more research into performance metrics for un-ranked subjective multi-label tasks is needed to draw a definitive conclusion.

7 Conclusions and Recommendations

This chapter will first highlight the main conclusions that were drawn from this research. Furthermore, the recommendations section adds onto the conclusion with recommendations for future research.

7.1 Conclusions

Single-value prediction

The study investigated how LLMs can detect subjective values in public discourse, focusing on different prompting methods. Chain-of-Thought prompting showed the best performance among tested methods, though none achieved high scores overall (see chapter 6). Considering practicality, the few-shot prompting method may be preferable due to its shorter run-time, despite slightly lower performance. None of the methods managed an F1 score of over 0.6, which indicates poor performance overall. When only considering these scores, one might conclude that LLMs are not (yet) equipped to accurately predict subjective underlying human values in argument statements in public discourse. However, as discussed in section 6.2, a soft-loss function approach for the multi-value predictions might yield different results [21].

Lack of suitable metrics

Another conclusion drawn from this research is that there is a lack of suitable metrics for evaluating the performance of LLMs on highly subjective tasks such as value detection. As can be seen in Table 1, multi-value scores are significantly lower than single-value majority label scores. Despite the LLM predictions appearing very reasonable on subjective observation, no metric currently represents this adequately. Ranking comparisons, such as Kemeny-Young[22], are not entirely suitable for task with this high level of subjectivity. As mentioned in section 6.1, these comparisons require a definitive ordered ranking and an equal amount of ranked items.

Capturing human subjectivity

Per-annotator prediction was done to investigate if an LLM is able to replicate a specific human's subjectivity. As can be seen in Figure 4, the scores for these predictions did not reach 0.5 in any of the runs. Another important thing to note in this experiment is the great inconsistency in scores between different runs for the same annotator. The combined consideration of both the low scores and the high inconsistencies in performance leads to the conclusion that LLMs are not very suitable for capturing human subjectivity.

7.2 Future research

As mentioned in chapter 6, the complexity of this task partly lies in the difficulty of comparing the multi-value predictions. 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 [21]. As mentioned in section 6.1, existing multi-value comparison metrics such as ranking comparisons are not very suitable for tasks with such a highly subjective nature. 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.

A secondary recommendation is to collect a more extensive set of annotations for this task. The fifty data points on a Frysian municipal energy policy used for evaluation in this research are insufficient to be considered representative of all public discourse. The sample size is limited, and the diversity of the topics of the argument is also lacking.

Lastly, a user study on people's ability to distinguish between LLM predictions and human annotations is recommended. After stating the difficulty of capturing the subjectivity of this task through objective metrics, it would be most interesting to see how other humans would evaluate the LLM's predictions in a blind user study. The main question being whether they would be able to tell the difference if all they were given was the argument statement and the annotated values.

8 Limitations

This section covers some of the limitations of this research. It dives into the quality and quantity of the data, the struggle to find metrics suitable for this task as well as the limitation of the Fleiss Kappa statistic for mapping annotator consensus.

8.1 Lack of data

The fifty data points on a Frysian municipal energy policy used for evaluation in this research are insufficient to be considered representative of all public discourse. The sample size is limited, and the diversity of the topics of the argument is also lacking. Consequently, many of the possible values were never assigned and therefore underrepresented in this study.

8.2 Lack of suitable metrics

The main conclusion of this research is that current evaluation metrics for machine learning tasks are insufficient for evaluation of tasks with such subjective nature. The current results contain metrics such as precision, recall and F1 scores, but these scores do not paint a complete picture of how well the values predicted by the model might also align with the given argument statement.

8.3 Limitation of Fleiss' Kappa

Currently, the only metric that is able to capture the degree of subjectivity of this task is the Fleiss Kappa statistic. However, relying on a single agreement coefficient can obscure complex patterns in annotation, such as diversity in underlying data, label similarities, varying difficulty of individual items and personal differences between annotators [17].

9 Responsible Research

This section aims to address the efforts made to ensure the integrity of this research. The two main components of responsible research that were most applicable to this researched were the use of the Dataset and ensuring the reproducibility of the experiment.

9.1 Data

The dataset that was provided to us contained residents' opinions on future municipal policies. The study from which this dataset was obtained had previously already anonymized the data, meaning that there was no personal data linked to the argument statements.

Even thought the data was not sensitive in that it contained personal information, it did contain information that the municipality of Súdwest-Fryslân did not want made public. For that reason no online version of LLMs were used; all experiments were run locally with the open-source LLama3 model.

There are some issues regarding the quality of the dataset that need to be addressed. First and foremost the quality of the data. The argument statements were peoples' typed additions to a ranking they have given that we did not have access to. This meant that the argument statements sometimes referred to unseen answers previously given, as well as them often being grammatically incorrect due to their typed nature. In addition to this, the quality of the automatic translations from Dutch to English was also adequate at most. This difference was only noticed by 2 out of 5 people from our project group, as we were the only dutch speaking annotators. This surely influence the tone of the arguments statement, and in an attempt to deal with these differences us Dutch annotators only looked at the English versions of the statements to ensure we had the same data points.

Lastly, a note on the annotation process: Since this task is so highly subjective, it is important to realize that these annotations are also highly subjective, and by no means an objective ground truth. One annotator even said "If I were to do this again tomorrow, the values would be different". This is also in part due to the large number of possible labels. Trying to choose out of 54 possible values was a challenging task.

9.2 Reproducibility

Sadly the data used for this research can not be made public. However, all code used in this experiment will be published online, and the LLM used is also freely available online [18].

References

[1] E. Blacksher, A. Diebel, P. Forest, S. D. Goold, and J. Abelson, "What Is Public Deliberation?," *Hastings Center Report*, vol. 42, pp. 14–16, Mar. 2012.

[2] S. H. Schwartz, "Basic Human Values," 2012.

[3] 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, Dec. 2021.

[4] 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), pp. 4459–4471, Association for Computational Linguistics, May 2022.

[5] R. Scollon, *Analyzing Public Discourse: Discourse Analysis in the Making of Public Policy*. Routledge, Oct. 2012. Google-Books-ID: sK53CGJJ72QC.

[6] S. Freeman, "Deliberative Democracy: A Sympathetic Comment," *Philosophy & Public Affairs*, vol. 29, no. 4, pp. 371–418, 2000. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1088-4963.2000.00371.x.

[7] G. Sayre-McCord, *Essays on Moral Realism*. Cornell University Press, 1988.

[8] S. H. Schwartz, "Are There Universal Aspects in the Structure and Contents of Human Values?," *Journal of Social Issues*, vol. 50, no. 4, pp. 19–45, 1994. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-4560.1994.tb01196.x.

[9] M. Weber and M. Reichardt, "Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models," Dec. 2023. arXiv:2401.00284 [cs].

[10] E. Liscio, A. Dondera, A. Geadau, C. Jonker, and P. Murukannaiah, "Cross-Domain Classification of Moral Values," in *Findings of the Association for Computational Linguistics: NAACL 2022* (M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, eds.), (Seattle, United States), pp. 2727–2745, Association for Computational Linguistics, July 2022.

[11] E. Liscio, "Axies: Identifying and Evaluating Context-Specific Values," 2021.

[12] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, O. Dirilen-Gumus, and M. Konty, "Refining the theory of basic individual values," *Journal of Personality and Social Psychology*, vol. 103, no. 4, pp. 663–688, 2012. Place: US Publisher: American Psychological Association.

[13] "Energy in Súdwest-Fryslân."

[14] M. Kiedrowicz and J. Stanik, "Selected aspects of risk management in respect of security of the document life-cycle management system with multiple levels of sensitivity.," pp. 231–249, Jan. 2015.

[15] T. R. Nichols, P. M. Wisner, G. Cripe, and L. Gulabchand, "Putting the Kappa Statistic to Use," *The Quality Assurance Journal*, vol. 13, no. 3-4, pp. 57–61, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qaj.481.

[16] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. Publisher: [Wiley, International Biometric Society].

[17] R. Artstein, "Inter-annotator Agreement," in *Handbook of Linguistic Annotation* (N. Ide and J. Pustejovsky, eds.), pp. 297–313, Dordrecht: Springer Netherlands, 2017.

[18] Ollama, "Llama3," 2024. https://ollama.com/library/llama3.

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," 2022.

[21] 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.

[22] B. Lausen, D. Van Den Poel, and A. Ultsch, eds., *Algorithms from and for Nature and Life: Classification and Data Analysis*. Studies in Classification, Data Analysis, and Knowledge Organization, Cham: Springer International Publishing, 2013.