Leveraging LLMs for Classifying Subjective Topics Behind Public Discourse

Ana Cristiana Marcu[1]

Supervisors: Luciano Cavalcante Siebert[1]**, Amir Homayounirad**[1]**, Enrico Liscio**[1]

[1]**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: Ana Cristiana Marcu
Final project course: CSE3000 Research Project
Thesis committee: Luciano Cavalcante Siebert, Amir Homayounirad, Enrico Liscio, Jie Yang

Abstract

Public deliberations play a crucial role in democratic systems. However, the unstructured nature of deliberations leads to challenges for moderators to analyze the large volume of data produced. This paper aims to solve this challenge by automatically identifying **subjective topics** behind public discourse by leveraging Large Language Models (**LLMs**). The study is structured around two core objectives: **Identifying Gold Labels** and **Exploring Subjective Human Labels**. The results highlight that fine-tuning the LLaMa-2 model with QLoRa outperforms other methods for Identifying Gold Labels, while the Few-Shot Chain of Thoughts method, enhanced with EmotionPrompt, is particularly effective in capturing subjective variations in human annotations. However, the study also underscores significant limitations, such as the dependency on large, high-quality annotated datasets and the tendency of models to produce hallucinations. These findings highlight the potential of LLMs to identify subjective topics behind public discourse, while also emphasizing the need for further research to address these challenges.

1 Introduction

Nowadays, due to an urgent demand for citizen participation in deliberations [1], engagement in public debates has gained greater significance. Deliberative democracy is now a thriving area and a vital component of the democratic systems [2]. Still, the unstructured nature of deliberations often leads to challenges for moderators to comprehend and analyze the large volume of data produced [3]. The problem arises when discussing scaling up deliberations because deliberative practices need to be sustained and effective on a larger scale, influencing broader decision-making processes and societal actions [4]. An essential first step in structuring deliberations is to identify the topics within them, as public debates often present various topics that uncover numerous hidden subtopics within the context. This approach aims to help moderators to understand the dominant contextual patterns in public discourse.

In every classification problem, labeled data is essential, but it necessitates employing a group of annotators to label the dataset, a process that is both costly and time-consuming. The challenge is further compounded when considering the subjectivity of annotators, particularly in tasks like topic identification [5]. Here, variations in interpretation, context understanding, text ambiguity, and personal biases can result in diverse annotations. Based on the hypothesis that disagreement should be treated as signal rather than noise, and minority voices should not be excluded [6, 7, 8], this research is divided into two parts: ***Identifying Gold Labels***, a single true set of possible labels, and ***Exploring Subjective Human Labels*** to discover each annotator's subjective decisions.

In recent years, the field of natural language processing (NLP) has undergone a significant transformation with the emergence of large language models (LLMs). The progress in this domain is largely attributed to increasing the scale of language models: enhancing the computational power, expanding the number of model parameters, and extending the training datasets [9]. Building on this foundation, these developments have led LLMs to approach human-level performance when it comes to processing and generating text with coherent communication [10], and performing complex tasks across diverse fields [11, 12]. Advanced LLMs offer a promising opportunity to revolutionize the identification of subjective data annotation.

Previous studies have extensively explored the capabilities of LLMs in diverse NLP applications, such as text classification [13, 14], topic modeling [15], as well as their capabilities to outperform humans in annotating datasets [16, 17, 18]. While the performance of LLMs for text annotation is promising, several aspects remain unclear and require further research, especially in leveraging LLMs to discover subjective topics in annotators' decisions during the labeling process. To fill this gap, this research aims to answer the main question:

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

Addressing this research question could redefine the landscape of subjective data annotation, ultimately contributing to the underlying problem of helping moderators structure deliberations. To help answer this main research question, two secondary questions have been formulated:

- *"What methods can we use to train LLMs to find subjective topics behind opinions?"*
- *"How can we evaluate the subjectivity of the topics generated by LLMs methods?"*

Addressing these questions will provide a clearer understanding of the process of utilizing LLMs as subjective annotators, the techniques used to identify subjective topics, and which method performs best. Moreover, to expand the possibilities for automatic data annotation in subjective contexts, the process of data preparation will be presented, from identifying topic labels in an unlabeled dataset to the manual annotation and aggregation of labels for evaluation. The manual annotation process will further provide insights on the level of subjectivity of the task, as complete label agreement towards a given set of items happens very rarely [19, 20, 21].

This research is structured as follows. Section 2 provides the background of the study, discussing topic modeling techniques and the role of subjectivity in identifying topics. Section 3 gives an overview of the data used during the experiments, detailing the label extraction, annotation, and aggregation process. Section 4 introduces the research methodology, detailing the experimental setup for prompting methods and fine-tuning the LLM. Section 5 presents the approach used to evaluate the methods and the results obtained from these experiments. A detailed discussion in Section 5 expands on the findings and addresses the study's limitations. The paper continues with Section 6, which considers the responsible research aspects implemented throughout the study. Finally, Section 7 summarizes the key findings, points out opportunities for further work, and concludes the research.

2 Background

This section explores topic modeling techniques to identify the main topics within a dataset and examines the role of subjectivity when identifying topics within a corpus of text. Thus, it sets the foundation for the methodologies and experiments discussed later in the research.

2.1 Topics modeling techniques

A topic is a central theme or subject that emerges from a text. Topics discussed during deliberations often possess a complex nature; they do not fit neatly within single disciplinary boundaries and may span multiple fields of expertise [22]. While there may be a clear main subject in any discussion, the context frequently reveals various subtopics that can be significant for a deeper understanding and analysis of extensive datasets.

To efficiently extract features from a vast corpus of text data, various text mining approaches have been developed with topic modeling being the most commonly employed technique [23]. Topic modeling is fundamentally an unsupervised approach that identifies underlying topics—represented as distributions of words—across a document corpus and determines each document's affinity to these topics. It is a technique to identify thematic structures within textual data, enabling efficient exploration of large textual datasets.

Prior work has been focusing on topic modeling techniques like Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Top2Vec, and BERTopic [24]. These techniques are forms of statistical modeling used in machine learning and NLP that identify hidden topical patterns within a collection of texts [25]. Topic modeling is particularly valuable in this research context, as it serves as an effective method for identifying dataset labels.

A comparative analysis of four methods determined the most suitable topic modeling technique [26]. The methods compared included LDA and NMF. LDA identifies topic patterns by assuming each document is a mixture of latent topics, each defined by a specific word distribution [27]. NMF reveals hidden patterns by factoring a matrix into two lower-ranking matrices. However, both methods overlook semantic relationships among words.

Given these limitations, two other methods were considered: Top2Vec and BERTopic. Top2Vec uses a pre-trained embedding model to map documents to a topic space but often produces overlapping topics, diluting clarity. In contrast, BERTopic leverages a pre-trained BERT model to generate nuanced topic representations, capturing semantic similarities without topic overlap.

2.2 Role of subjectivity in identifying topics

When addressing classification problems, especially in the realm of machine learning, researchers usually first determine the true labels for their study [16]. This requires annotators to label data, a process that can lead to varying results among annotators, particularly in NLP tasks [28, 21]. The disagreement between annotators is linked to the subjectivity of the task. Subjectivity arises from an individual's unique perceptions, biases, opinions, and personal experiences, all of which shape their understanding of the world and are not entirely objective. Additionally, external factors such as the time of day can influence how a person annotates, potentially affecting subjective tasks. The same individual may annotate differently depending on when they perform the task.

Despite efforts to minimize inter-annotator disagreement, every project involving linguistic annotation or text annotation tasks has to deal with cases of diverging perceptions among annotators [5]. Based on the hypothesis that a single true label does not exist for subjective tasks, researchers suggested that disagreement has to be treated as signal and not noise. Thus, it has been proposed to use *human label variation* rather than the term *disagreement* [8] to capture that more views might be plausible.

3 Data

The dataset employed in this study is not publicly accessible and covers a deliberation on future energy policies, specifically concerning the Energy in Súdwest-Fryslân[1]. The deliberation involved 1376 local residents, focusing on understanding their perspectives on shaping future energy policies for their municipality. The dataset contains 482 responses from the deliberation, translated into three languages: Frisian, Dutch, and English. This research paper focuses on the English version of their replies.

3.1 Topic modeling with BERTopic

The dataset provided was unlabeled, with no predefined labels for the topics. Consequently, the first step was to identify the possible labels. Based on the comparative analysis shown in the study previously mentioned in Section 2.1, BERTopic [29] was selected to generate the topic labels. The unsupervised approach operates through three primary phases, as seen in Figure 1. Using a pre-trained BERT model, each document is embedded into a high-dimensional vector space, which is then reduced using UMAP in order to cluster the embeddings using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm, which is effective in identifying clusters of varying density.

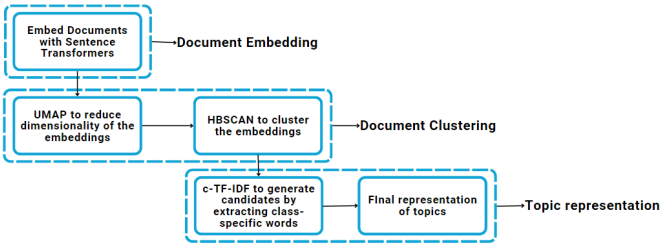

Figure 1: Overview of BERTopic's primary phases: Document Embedding, Document Clustering and Topic representation

Finally, for each cluster, a topic representation is derived using the class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) algorithm. Unlike traditional

[1] https://www.tudelft.nl/en/tpm/pve/case-studies/energy-in-sudwest-fryslan

TF-IDF, which measures the importance of words within individual documents, c-TF-IDF evaluates the significance of words for each cluster of documents, thereby enhancing the clarity and coherence of the identified topics [29]. This is done by computing the weight of every word within a cluster and taking the most significant ones as follows. For a term x within cluster c:

$$w_{x,c} = ||tf_{x,c}|| \cdot \log\left(1 + \frac{A}{f_x}\right) \quad (1)$$

where $tf_{x,c}$ represents the frequency of word x in cluster c, f_x is the frequency of word x across all clusters, and A is the average number of words per cluster.

The final labels were determined after merging similar topics and manually defining the names of the labels based on the results of BERTopic (Appendix A). Data items that were not clustered into one of the six topics remained unclassified or less strongly associated, hence not fully accounted for in the primary topics. The labels, as well as each cluster size, can be seen in Table 1.

Label No.	Topic Label	Cluster Size
1	Municipality and residents engagement in the energy sector	133
2	Energy storage and supplying energy in The Netherlands	88
3	Wind and solar energy	44
4	Market Determination Dynamics	41
5	Landscapes and windmills tourism	22
6	Hydrogen energy pipeline networks	14
7	Unclassified/Outliers	140

Table 1: Overview of Topics Identified by BERTopic and Cluster Sizes

3.2 Data annotation and aggregation

Data annotation is the process of assigning labels to individual data items, in this case, assigning topics to text. Disagreement among annotators is a common challenge in NLP [30], often attributed to the subjective interpretations of the text by each annotator. Consequently, the data annotation process is crucial for exploring the subjectivity in identifying topics. The annotation team comprised five computer science students familiar with the dataset. Each annotator was briefed on the multi-label classification problem and annotated 50 data items, influenced by their personal opinions, contextual understanding, biases, and external factors.

To further investigate the subjectivity of the task, the Fleiss' Kappa metric [31] was employed to assess the agreement between annotators for each label. This metric is especially useful in this case because, unlike Cohen's Kappa, Fleiss' Kappa can evaluate agreement for any number of raters. It is computed by dividing the degree of agreement actually achieved above chance by the degree of agreement that is attainable above chance. This relationship is formally represented by the equation below:

$$\kappa = \frac{\hat{P} - \hat{P}_e}{1 - \hat{P}_e} \quad (2)$$

The results of the annotation process are summarized in Table 2, which displays the Fleiss Kappa scores for each topic, as well as their interpretation [32]. These scores provide insight into the degree of consensus among the annotators regarding the classification of each topic. Based on the interpretation of the results, it can be concluded that there is generally moderate agreement among annotators. However, the subjectivity of annotations is evident from the lack of consensus among them.

Label No.	Total Number of Annotations per Label	Fleiss Kappa Score	Interpretation
1	132	0.65	Substantial Agreement
2	66	0.5	Moderate Agreement
3	48	0.48	Moderate Agreement
4	44	0.34	Fair Agreement
5	34	0.6	Moderate Agreement
6	0	NaN	No significant result

Table 2: Fleiss Kappa Scores for Inter-annotator Agreement and Total Annotations per Label by 5 Raters

The data annotation procedure was followed by data aggregation, used for the training process and evaluation. A majority vote (>50%) was chosen as the aggregation method for *Identifying Gold Labels*. In addition, a no-aggregation strategy was pursued, to investigate how well LLMs can detect the annotations for each rater, thereby analyzing the detection of subjectivity for each annotator and *Exploring Subjective Human Labels*.

One important consideration is the number of data items annotated as well as the balance of labels. Annotators labeled 50 data items, with a preference for the first topic label, as can be observed in Table 2. By connecting the results to the distribution of topics discovered by BERTopic, as outlined in Table 1, it is observed a generally strong alignment between the topics identified by the model and the frequency of annotations by the raters. Labels such as 1 and 2 not only feature prominently in the dataset but also receive a correspondingly high number of annotations, indicating their clear definition and relevance in the dataset. Conversely, the lack of annotations for 6 despite its presence, suggests that this topic might not be present in the selection of data items.

4 Methodology

The research aims to identify subjective topics behind public discourse by leveraging LLMs. To achieve this, the problem is defined as a *multi-label classifier*. Formally, the goal can be defined as follows. Let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n texts and $L = \{l_1, l_2, \ldots, l_k\}$ a set of k binary labels representing the possible topics from the dataset. Each text $t_i \in T$ can be associated with a subset of labels from L. Finally, a function is defined $f : T \to \{0, 1\}^k$ where $f(t_i) = (y_{i1}, y_{i2}, \ldots, y_{ik})$ and $y_{ij} \in \{0, 1\}$ indicates the presence (1) or absence (0) of

label l_j for text t_i. The goal of the problem is to use LLM prompting and training methods $\hat{f}$ that map each text t_i to a vector of predicted labels $\hat{f}(t_i) = (\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{ik})$ such that $\hat{y}_{ij}$ is as close as possible to the true label y_{ij}.

Subjectivity is encapsulated in each annotator's labeling decisions, which serve as individual true values for *Exploring Subjective Human Labels*. The labels can be aggregated for *Identifying Gold Labels* to create the true values for the multi-label classification problem defined above. This section aims to explain the experiments conducted to address the identification of subjective topics. It covers the choice of the LLM used, prompt engineering strategies employed, and the process of fine-tuning the LLM.

4.1 LLaMa-2 model

The LLaMA-2 7B model was selected as the LLM for this research project, Meta's foundational open-source model with 7B parameters that uses an optimized transformer architecture. LLaMa models series tries to democratize the access and study of LLMs, as they run on a single GPU [33] with the support of Ollama tool[2] while being famous for parameter efficiency and instruction tuning. Additionally, as an open-source LLM, it provides significant data protection benefits, ensuring that data is not shared with third parties, thereby enhancing security and confidentiality [34].

Another important consideration was the temperature parameter of the LLM. The temperature parameter regulates the amount of randomness, leading to more diverse outputs. To explore the subjectivity of the task, a temperature of 0.7 was chosen to explore more creative outputs [35].

4.2 Prompt engineering strategies

Prompt engineering strategies were employed during the experiments to compare which one performs the most accurately for *Identifying Gold Labels* (using majority vote labels for training and evaluation) and *Exploring Subjective Human Labels* (using individual annotator labels for training and evaluation). The specific prompting techniques used for each task are summarized in Table 3 and the overview of the prompt structure can be seen in Figure 2.

Given the recent advancements in LLMs, prompt engineering has become increasingly significant, so that the strategic design of the inputs could provide more accurate outputs. It serves as a bridge between human communication and the computational capabilities of LLMs, by providing contextual input and specifying desired output formats. The prompt structure and the diagrams of the prompting techniques used during the experiments are detailed in Appendix B, C and D.

Zero-shot prompting means that the prompt contains only the task that the LLM is supposed to do and does not contain any examples. This method leverages the LLM's generalization capabilities. However, due to its nature, it cannot account for the subjectivity of each annotator and is thus used solely for classifying text into a single true set of topics.

[2]https://ollama.com/

Label category	Prompting Techniques
Identifying Gold Labels	Zero-Shot Prompting; Chain of Thoughts with Zero-Shot Prompting; Few-Shot Prompting; Chain of Thoughts with Few-Shot Prompting
Exploring Subjective Human Labels	Few-Shot Prompting; Chain of Thoughts with Few-Shot Prompting; Two-step Rephrase and Respond + Few-Shot Chain of Thoughts Prompting; Few-Shot Chain of Thoughts Prompting + EmotionPrompt; Two-step Rephrase and Respond + Few-Shot Chain of Thoughts Prompting

Table 3: Prompt engineering strategies used for Identifying Gold Labels and Exploring Subjective Human Labels

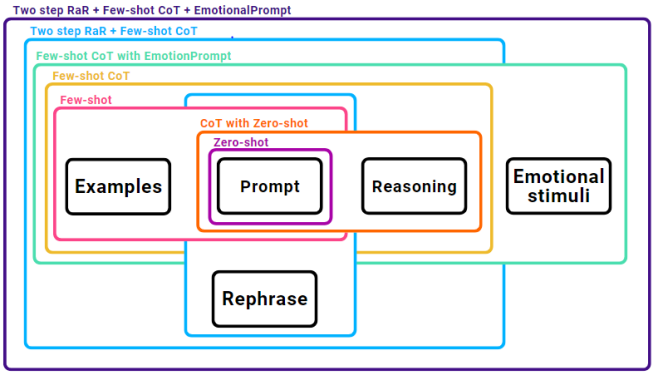

Figure 2: Overview of Prompt Engineering Strategies used with LLaMa-2 7B for Identifying Gold Labels and Exploring Subjective Human Labels

Zero-shot Chain of Thoughts prompting uses prompting twice, to extract both reasoning and the final answer. The first prompt extracts the reasoning path from the language model by adding `"Let's think step by step"` at the end of the prompt and then the second prompt is used to extract the answer in the correct format from the reasoning text by adding `"Therefore, the final answer is "` [36].

Few-shot prompting was adopted to identify both gold labels and the subjectivity of the annotators. In this approach, the model is given a few examples of the task during inference as conditioning, but the model's weights are not updated. The few-shot method operates by providing K examples of context and completion, followed by one final example of context, with the model expected to provide the completion [37]. For this research, $K = 5$, with the note that for *Exploring Subjective Human Labels*, the examples included each annotator's labels to identify the subjective views.

Few-shot Chain of Thoughts prompting is a method that enhances the Few-shot prompting by adopting a logical and sequential approach that enables more scientific and human-like reasoning in LLMs [38]. This method is particularly useful when obtaining each annotator's labels, as it helps the LLM recognize the subjectivity inherent in the task. For this, two reasoning strategies were adopted. The first strategy involved adding the phrase `"Let's think step by step"` at the end the prompt, primarily used to assist in *Identifying Gold Labels* effectively. The second strategy utilized a comprehensive prompt designed to delve into the subjective reasoning behind each annotator's labeling decisions: `"For each label you assign, please provide a detailed explanation of your reasoning. Explain why each annotator assigned each topic to the text. Remember that they are classifying text into topics. We aim to capture the subjective decisions of each annotator in labeling the data based on their previous labeling decisions and to find correlations from their past choices."`

Two-step Rephrase and Respond + Few-shot Chain of Thoughts prompting is a method that aims to reduce the misinterpretation of questions by LLMs. Rephrase and Respond (RaR) method [39] prompts the LLM to articulate the question and then respond in a two-step process: `"Given the above task, rephrase and expand it to help you do better answering. Maintain all information in the original question."`. The research explores this additional step, which complements the Chain of Thoughts (CoT) method, to achieve performance improvements.

Few-shot Chain of Thoughts prompting + EmotionPrompt leverages the fact that previous studies in psychology have demonstrated that adding emotional stimuli related to expectancy and confidence can positively impact individuals and improve the responses of LLMs [40]. EmotionPrompt is a type of prompt that leverages the emotional intelligence of LLMs. In conjunction with providing examples and reasoning, a social cognitive stimulus was introduced in the prompt: `"Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results."`

Two-step Rephrase and Respond + Few-shot Chain of Thoughts prompting + EmotionPrompt is an additional proposed method that combines the elimination of question misinterpretation by the LLM, reasoning capabilities over the subjective patterns of the annotators, and emotional stimuli. This approach aims to integrate the previously discovered prompt engineering methods to determine if it outperforms the earlier presented methods.

4.3 Fine-tuning LLaMa-2 with QLora

Full parameter fine-tuning involves adjusting all the parameters across all layers of a pre-trained model. While this approach often results in optimal performance, it requires significant computational resources. In contrast, Parameter Efficient Fine Tuning (PEFT) allows for model refinement using minimal resources. QLoRA incorporates Quantization and Low-Rank Adapters (LoRA) to make the process more accessible, as can be seen in Figure 3 [41]. This approach democratizes fine-tuning by enabling the optimization of extremely large models, which have billions of parameters, using relatively small, highly available GPUs [41].

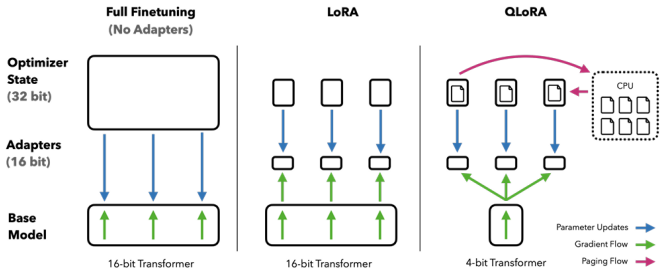

Figure 3: Fine-tuning methods: Full Fine-tuning with no adapters, LoRa and QLoRa [41]

Due to the low number of annotated data, a weakly supervised learning approach was implemented [42]. It combines the small amount of human-labeled data with a large number of weak labels generated by the BERTopic algorithm. Originally used to identify dataset labels, the algorithm now assigns each text cluster to a topic, thereby generating these weak labels. As this cannot cover the subjectivity of each annotator, this method was only used for *Identifying Gold Labels*. Therefore, all 482 data items were used in the fine-tuning process distributed across training, validation, and test sets in an 80/10/10 split. The data annotated by the annotators was evenly divided between the train and test set with a 50/50 ratio.

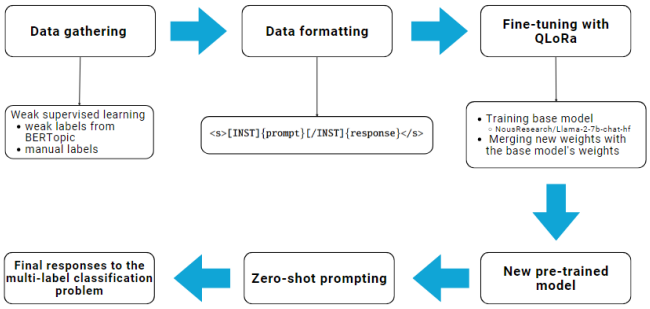

Figure 4: Topic Extraction Workflow: From Data Gathering to Prompting the Fine-Tuned Model

After gathering the data, each prompt-responses pair was formatted in an instruction-based manner:

`<s>[INST]{prompt}[/INST]{response}</s>`

The formatted data was then ready for the training of the base model: *NousResearch/Llama-2-7b-chat-hf* with an eval-

uation step of 20. The training was limited to two epochs to avoid overfitting, as extending beyond this resulted in the model fitting too closely to the training data. After the training, the new weights were merged with the base model's weights, resulting in a new pre-trained model. This new model was then used for *Identifying Gold Labels* using a Zero-shot prompting strategy, without including any additional context. An overview of the process of extracting the topics using a fine-tuned model can be seen in Figure 4.

5 Results

The results of this study provide insights into the accuracy of different methods implemented to optimize the performance of LLaMa-2 in *Identifying Gold Labels* and *Exploring Subjective Human Labels*. By comparing various approaches, including Zero-shot, Few-shot, and fine-tuning techniques, the research highlights which method is more suitable for each task. This section details the quantitative results from these methods, evaluated using the Micro F1-score, and describes the preprocessing steps performed prior to evaluation.

5.1 Preproccesing results and evaluation

To evaluate the methods, each training method was run ten times for the first 50 data items, and the results were aggregated using a majority vote (>50%). An exception was made for the fine-tuning method, which used a subset of 25 data items from the labeled data by annotators and another 23 from the weakly labeled data.

Due to the unpredictability of the LLM's output [43, 44], especially for methods that involved reasoning, a preproccesing step was performed. This involved eliminating any extra information provided by the LLM and ensuring that the output was formatted as follows: {*'Topics': Topic1, Topic2,* ...}. The primary issue was that the output list of topics often did not match the topics from the label list (see Table 1). Instead, it contained contextually similar topics but with different phrasing. Therefore, a mapping between the topics from the output list and the label list was done using embeddings, specifically Sentence Transformers[3]. To determine whether a topic is mapped to one of the topics from the original list, cosine similarity was used with a threshold of 0.7. This ensured that no relevant values were discarded from the results.

After the preprocessing of results was done, the training methods can be evaluated. A suitable choice for an evaluation metric was subject to the class imbalance in our dataset. Therefore, a Micro Averaged F1-Score was chosen to take into account the unbalance in the dataset:

$$F1_{\text{micro}} = \frac{2 \cdot Precision_{\text{micro}} \cdot Recall_{\text{micro}}}{Precision_{\text{micro}} + Recall_{\text{micro}}} \quad (3)$$

$$Precision_{\text{micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \quad (4)$$

$$Recall_{\text{micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \quad (5)$$

[3]https://huggingface.co/sentence-transformers

5.2 Results for Identifying Gold Labels

After preprocessing and calculating the Micro F1-score, the results were derived as follows. The outcomes for *Identifying Gold Labels*, with training and evaluation performed using the majority vote aggregated results, are presented in Table 5.

The Zero-shot approach served as a baseline with a score of 0.64 while incorporating CoT reasoning slightly improved the score to 0.657. The Few-shot and Few-shot with CoT methods both significantly enhanced performance, achieving a Micro F1- score of 0.817. The fine-tuning method achieved the highest score, a Micro F1-score of 0.865. These findings suggest that while CoT reasoning can enhance model performance in a Zero-shot context, substantial improvements are predominantly achieved through Few-shot prompting and fine-tuning.

5.3 Results for Exploring Subjective Human Labels

The results for *Exploring Subjective Human Labels* for each prompting method per annotator are shown in Table 7, and the averaged results of the annotators per method are displayed in Table 6 for better comparison.

The Few-shot CoT v2 + EmotionPrompt method achieved the highest averaged Micro F1-Score among all the methods tested, with a score of 0.7824. This indicates that incorporating emotional stimuli into the Few-shot CoT v2 significantly enhances performance in classifying subjective topics. Moreover, the Few-shot CoT method scores variably (0.715 and 0.75 for v1 and v2, respectively), suggesting iterative improvements from the first prompt version to the second one. The RaR + Few-shot CoT v2 method achieved the lowest score among all the methods tested, with a score of 0.682, indicating that rephrasing the prompt may not positively contribute to performance in this particular use case.

Comparing RaR + Few-shot CoT v2 + EmotionPrompt (0.7824) to Few-shot CoT v2 + Emotion-Prompt (0.779) suggests that while the addition of EmotionPrompt generally enhances performance, its effectiveness might depend on the foundational method it is paired with. The lower performance of the RaR + Few-shot CoT v2 method without the EmotionPrompt and the better performance of Few-shot CoT v2, when paired with EmotionPrompt, indicate that the addition of RaR may not enhance, and may indeed detract from the overall accuracy. This implies that the positive contributions are primarily from the EmotionPrompt.

5.4 Effectiveness of Fine-tuning with QLora

To assess the effectiveness of fine-tuning LLaMa-2 with QLora, the loss for both the training and validation sets was monitored. The training loss converged to 0.62, while the evaluation loss converged to 0.85, as seen in Figure 5. The results were gathered after two epochs. Using more epochs led to an increase in validation loss while the training loss continued to decrease, indicating potential overfitting.

6 Discussion and limitations

The study's evaluation of various training methods for LLMs to classify subjective topics behind public discourse presents

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	0.865

Table 5: 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	0.7824
RaR + Few-shot CoT v2 + Emotion-Prompt	0.779

Table 6: Averaged Micro-F1 Score Results for Prompting Methods for Exploring Subjective Human Labels

Prompting Method	Annotator1	Annotator2	Annotator3	Annotator4	Annotator5
Few-shot	0.767	0.637	0.756	0.855	0.767
Few-shot CoT	0.733	0.593	0.715	0.785	0.748
Few-shot CoT v2	0.75	0.715	0.763	0.756	0.77
RaR + Few-shot CoT	0.725	0.615	0.707	0.711	0.652
Few-shot CoT + EmotionPrompt	0.742	0.756	0.826	0.84	0.748
RaR + Few-shot CoT + EmotionPrompt	0.758	0.744	0.826	0.815	0.752

Table 7: Micro-F1 Score Results for Prompting Methods per each annotator for Exploring Subjective Human Labels

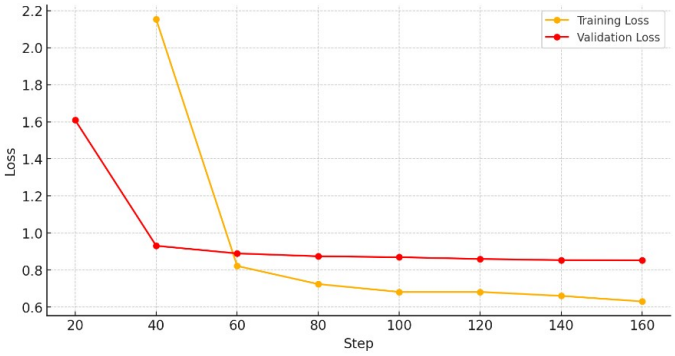

Figure 5: Training and Validation Loss over Steps for fine-tuning LLaMa-2 with QLora over two epochs

several notable findings and implications. The improvement in performance from Zero-shot methods to Few-shot and fine-tuning approaches highlights the LLMs' capability to adapt to complex classification tasks through enhanced prompting and training techniques. CoT reasoning, particularly, has been shown to add value by enabling the models to process and articulate their reasoning steps, thereby increasing the accuracy of classifications in contexts where logical reasoning is crucial.

The enhanced performance observed in *Identifying Gold Labels* compared to *Exploring Subjective Human Labels* could be attributed to the relatively objective nature of the task of identifying topics, as indicated by the moderate agreement among annotators. This suggests that variations between annotations might not be substantial. However, the model might expect variations in labeling patterns when given the task of identifying the labels for multiple annotators, but such variations were minimally present in this study. Additionally, the limited number of annotations made the process of fine-tuning the model to identify annotators' labeling patterns not possible for this study. Nevertheless, the improvement in performance through fine-tuning, suggests that further refinement of the models could potentially enhance their ability to capture subtler variations in annotation behavior.

This study relied on subjective annotations from five individuals with similar backgrounds. This is recognized as a limitation, as including a broader range of perspectives in the annotation process is essential to capture the variability in human judgments. To avoid excluding minority voices, a larger and more diverse pool of annotators would be necessary.

However, the study's results also bring to light the inherent limitations of LLMs, such as their tendency towards generating hallucinations—fabricated information not supported by the input data [45]. This erroneous reasoning can undermine the reliability of model outputs, especially in high-stakes settings such as public discourse analysis, where the accuracy and trustworthiness of information are paramount [46].

Another important limitation of the LLMs is the dependency on high-quality data. LLMs' usefulness in generating annotations depends on large, high-quality datasets. Nevertheless, this presents a scalability issue for LLM-based annotation projects. The research depended on manual annotations and was therefore limited to a small dataset. Consequently, the evaluation was carried out on a restricted number of annotations. This limited data size may hinder the generalizability of the findings. Moreover, this applies to the performance of the fine-tuning method, as models trained under weak supervision might not achieve the same level of accuracy as those trained with fully labeled, high-quality datasets.

7 Responsible Research

In the emerging field of artificial intelligence (AI) technologies, history has shown that new ethical dilemmas frequently arise. This research, which primarily utilizes LLMs, faces its unique set of challenges. As such, this section is dedicated to discussing the ethical considerations and reproducibility of the research. The aim is to mitigate any potential negative social impacts stemming from the research and to provide a foundation for academia to further explore and expand upon this classification task.

7.1 Ethical considerations

Engineering ethics establish professional standards for practice. To ensure this, reflecting on the impact of the methods developed is essential. A first important consideration is that the annotation process for this research was constrained by time limitations, resulting in a relatively small dataset for training and evaluating the methods. Despite these constraints, the findings suggest that LLMs hold promise for identifying subjective topics. Moreover, to uphold the standards of privacy and ensure unbiased data collection, all annotators involved in the project were assigned random numbers as identifiers, preserving their anonymity throughout the research process. The dataset used for the study was also anonymized by experts, leading to a negligible risk of potential reidentification.

Future work should aim to expand the dataset with additional annotations and test these techniques across more diverse datasets to verify the robustness and performance of the methods. As this research progresses, it is important to incorporate ethical considerations to ensure that the deployment of these models adheres to principles of fairness, transparency, and accountability. This approach is crucial as the research is still in the preliminary stages of exploring how LLMs can effectively handle variations in human labeling.

7.2 Reproducibility

Reproducibility, defined as the ability of an independent team to duplicate results using the same methods under the same conditions, is a pillar of scientific progress and lends credibility to research findings. This study faces several challenges in achieving reproducibility. Notably, the datasets used are private and restricted due to confidentiality, limiting external validation and potentially affecting the generalizability of the results. Additionally, the inherent variability in LLM performance suggests that our results might not consistently replicate across the same settings if the temperature parameter of the LLM is different than 0 (which is the case for this study). An attempt to mitigate this issue includes conducting experiments multiple times and utilizing a data aggregation method to stabilize the results, although variability remains a concern.

To address these and ensure reproducibility to the best extent possible, the study's codebase has been made available as open-source on GitHub[4]. This allows researchers to replicate the work, provided they have access to the necessary datasets, which can be obtained by contacting the authors of the case study[5]. Additionally, detailed structures of the prompts used are provided in the appendix of this document, offering transparency and aiding in the reproduction of the study's procedures.

These challenges are acknowledged, and there is an emphasis on the necessity for further studies to verify these findings under various conditions. This effort aligns with our goal to contribute to the dialogue on the capabilities and limitations of LLMs in identifying subjective topics behind public discourse.

8 Conclusions and Future Work

This research explored the potential of LLMs to classify subjective topics in public discourse, thereby assessing their suitability for text classification challenges. Identifying the topics within a public deliberation could lead to a sustainable and effective practice in comprehending and analyzing the large volumes of data produced during debates. Initially, the study outlined the process of extracting potential labels from the dataset by using the BERTopic method [29], followed by a detailed description of data annotation and aggregation procedures. Subsequently, the research examined prompting techniques and fine-tuning to determine the most effective method for subjective topic classification. Thereby, the research was divided into two principal components: *Identifying Gold Labels* and *Exploring Subjective Human Labels*.

For *Identifying Gold Labels*, the most effective method proved to be the fine-tuning of the LLaMa-2 model using the QLora [41], achieving a Micro F1-score of 0.865. This method utilized a weakly supervised learning technique [42], which leveraged BERTopic's weak labels to generate the annotated dataset. However, this approach requires a substantial pre-annotated dataset and carries the risk of overfitting, potentially diminishing the model's effectiveness on novel, unseen data. Despite these challenges, alternative methods like Few-shot and Few-shot CoT also demonstrated significant performance.

For *Exploring Subjective Human Labels*, the Few-shot CoT v2 enhanced with EmotionPrompt achieved the highest accuracy, with a Micro F1-score of 0.782. This method underscores the benefits of integrating emotional stimuli with Few-shot CoT. However, the study also highlights the limitations in current LLM applications - primarily their reliance on high-quality, well-annotated datasets and their susceptibility to generating unreliable reasoning or hallucinations, particularly when employing the CoT prompting method.

To further assess the reliability of the CoT methods, conducting a qualitative analysis of the reasoning behind the annotators' labeling decisions could enhance the understanding of the hallucination issue. Additionally, exploring strategies to prevent such hallucinations would not only improve the method's reliability but also validate its performance.

An interesting direction for future work is to empirically analyze the impact of the temperature parameter on LLMs for classification problems. This study used a temperature

[4]https://github.com/AnaCMarcu/ClassifyingSubjectiveTopics

[5]https://www.tudelft.nl/en/tpm/pve/case-studies/energy-in-sudwest-fryslan

setting of 0.7, but further exploration with different values could yield insightful results. Additionally, incorporating soft probabilistic labels during experiments could be beneficial, as they may provide a more nuanced understanding of annotator perceptions.

Future research should also focus on expanding the annotated dataset to enhance the training and evaluation of LLM training methods. A larger dataset would not only provide more robust evaluations but also support the development of a more accurately fine-tuned model for *Identifying Gold Labels*, eliminating the reliance on weak labels. This expansion would also facilitate the fine-tuning of models for *Exploring Subjective Human Labels*. To further enhance the potential of fine-tuned models, prompting techniques could be incorporated. Moreover, to avoid excluding minority voices and to ensure the detection of variability in human judgment, a larger and more diverse pool of annotators will be necessary.

Further exploration of different LLMs could provide insights into which models are most effective for this specific task, an area yet to be explored. Additionally, future studies should address the ethical implications of deploying LLMs in sensitive areas like public policy and discourse. It is vital to ensure that these powerful tools are used responsibly and contribute positively and equitably to societal discourse analysis.

References

[1] D. Schleifer and A. Diep, "Strengthening democracy: What do americans think," *New York, NY: The*, 2019.

[2] J. S. Fishkin, *Democracy and deliberation: New directions for democratic reform*. Yale University Press, 1991.

[3] 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.

[4] A. Bächtiger and A. Wegmann, "Scaling up deliberation," *Deliberative democracy: Issues and cases*, pp. 118–135, 2014.

[5] M. Sandri, E. Leonardelli, S. Tonelli, and E. Ježek, "Why don't you do it right? analysing annotators' disagreement in subjective tasks," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2428–2441.

[6] M. van der Meer, N. Falk, P. K. Murukannaiah, and E. Liscio, "Annotator-centric active learning for subjective nlp tasks," *arXiv preprint arXiv:2404.15720*, 2024.

[7] L. Aroyo and C. Welty, "Truth is a lie: Crowd truth and the seven myths of human annotation," *AI Magazine*, vol. 36, no. 1, pp. 15–24, 2015.

[8] B. Plank, "The'problem'of human label variation: On ground truth in data, modeling and evaluation," *arXiv preprint arXiv:2211.02570*, 2022.

[9] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[10] B. A. y Arcas, "Do large language models understand us?" *Daedalus*, vol. 151, no. 2, pp. 183–197, 2022.

[11] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[12] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[13] A. Peña, A. Morales, J. Fierrez, I. Serna, J. Ortega-Garcia, I. Puente, J. Cordova, and G. Cordova, "Leveraging large language models for topic classification in the domain of public affairs," in *International Conference on Document Analysis and Recognition*. Springer, 2023, pp. 20–33.

[14] N. Hagar, "Llm-generated labels for topic classification," Medium, 2023, accessed: 2023-05-06. [Online]. Available: https://nicholashagar.medium.com/llm-generated-labels-for-topic-classification-4891698bb529

[15] Y. Mu, C. Dong, K. Bontcheva, and X. Song, "Large language models offer an alternative to the traditional approach of topic modelling," *arXiv preprint arXiv:2403.16248*, 2024.

[16] Z. Tan, A. Beigi, S. Wang, R. Guo, A. Bhattacharjee, B. Jiang, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation: A survey," 2024.

[17] X. He, Z. Lin, Y. Gong, A. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen *et al.*, "Annollm: Making large language models to be better crowdsourced annotators," *arXiv preprint arXiv:2303.16854*, 2023.

[18] J. Li, "A comparative study on annotation quality of crowdsourcing and llm via label aggregation," *arXiv preprint arXiv:2401.09760*, 2024.

[19] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein, "The disagreement deconvolution: Bringing machine learning performance metrics in line with reality," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[20] R. Egger and J. Yu, "Identifying hidden semantic structures in instagram data: a topic modelling comparison," *Tourism Review*, vol. 77, no. 4, pp. 1234–1246, 2021.

[21] 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, 2021.

[22] P. Levine, A. Fung, and J. Gastil, "Future directions for public deliberation," *Journal of Deliberative Democracy*, vol. 1, no. 1, 2005.

[23] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, 2010, pp. 80–88.

[24] N. C. Albanese, "Topic modeling with lsa, plsa, lda, nmf, bertopic, top2vec: a comparison a comparison between different topic modeling strategies including practical python examples," *Towards Data Sci*, 2022.

[25] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: A comparative analysis," *Frontiers in artificial intelligence*, vol. 3, p. 42, 2020.

[26] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, p. 886498, 2022.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[28] P. Röttger, B. Vidgen, D. Hovy, and J. B. Pierrehumbert, "Two contrasting data annotation paradigms for subjective nlp tasks," *arXiv preprint arXiv:2112.07475*, 2021.

[29] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[30] 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.

[31] K. A. Hallgren, "Computing inter-rater reliability for observational data: an overview and tutorial," *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012.

[32] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[34] A. Spirling, "Why open-source generative ai models are an ethical way forward for science," *Nature*, vol. 616, no. 7957, pp. 413–413, 2023.

[35] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, "Is temperature the creativity parameter of large language models?" *arXiv preprint arXiv:2405.00492*, 2024.

[36] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[37] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[38] O. Fagbohun, R. M. Harrison, and A. Dereventsov, "An empirical categorization of prompting techniques for large language models: A practitioner's guide," *arXiv preprint arXiv:2402.14837*, 2024.

[39] Y. Deng, W. Zhang, Z. Chen, and Q. Gu, "Rephrase and respond: Let large language models ask better questions for themselves," *arXiv preprint arXiv:2311.04205*, 2023.

[40] 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.

[41] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[42] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[43] N. Potyka, Y. Zhu, Y. He, E. Kharlamov, and S. Staab, "Robust knowledge extraction from large language models using social choice theory," *arXiv preprint arXiv:2312.14877*, 2023.

[44] T. Hagendorff, "Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods," *arXiv preprint arXiv:2303.13988*, 2023.

[45] 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.

[46] B. Jiang, Z. Tan, A. Nirmal, and H. Liu, "Disinformation detection: An evolving challenge in the age of llms," in *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 2024, pp. 427–435.

A BERTopic results

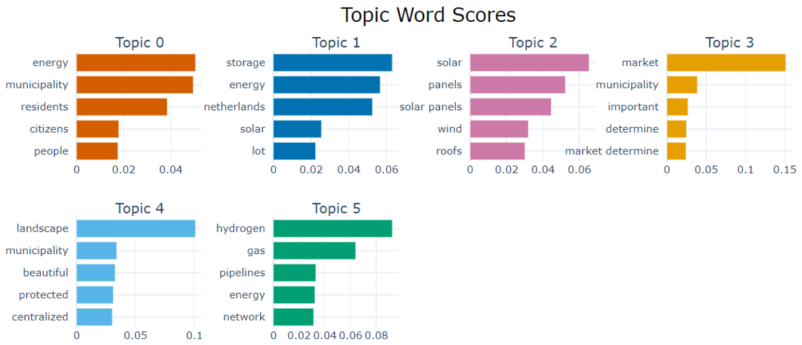

Figure 6: Results achieved after running the BERTopic algorithm

B Prompts content

Prompt parts	Content
system_prompt	You assign texts into topics. Answer with just the correct topics found in the text. Your answer should have maximum three topics based on the content of the text. Do not add any topics that are not listed. If no topics apply, respond with 'None of the topics are applicable'.
system_prompt_subjectivity	You assign texts into topics. For each annotator, identify up to three relevant topics for the text based on their previous responses. The topics should be relevant for the given text. Only use topics that have been listed; do not introduce new ones. If no topics apply to a text, respond with 'None of the topics are applicable'.
user_prompt(text)	**Text:** + (text) **Topics:** 1. Municipality and residents engagement in the energy sector 2. Energy storage and supplying energy in The Netherlands 3. Wind and solar energy 4. Market Determination Dynamics 5. Landscapes and windmills tourism 6. Hydrogen energy pipeline networks
prompt_format	The response should have the following format: {Topics: Topic1, Topic2, ...}
prompt_format_subjectivity	The response should have the following format: Annotator 1 - {Topics: Topic1, Topic2, ...} Annotator 2 - {Topics: Topic1, Topic2, ...} Annotator 3 - {Topics: Topic1, Topic2, ...} Annotator 4 - {Topics: Topic1, Topic2, ...} Annotator 5 - {Topics: Topic1, Topic2, ...}
reasoning_prompt	Let's think step by step.
reasoning_prompt_v2	For each label you assign, please provide a detailed explanation of your reasoning. Explain why each of the annotators assigned each topic to the text. Remember that they are classifying text into topics. We aim to capture the subjective decisions of each annotator in labeling the data based on their previous labeling decisions and to find correlations from their previous decisions.
rephrase_and_respond_prompt	Given the above task, rephrase and expand it to help you do better answering. Maintain all information in the original question.
emotion_prompt	Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results.

Table 4: Prompt content (the prompt part are used to explain the prompt structure in Appendix C and Appendix D)

C Prompts structure for Identifying Gold Labels

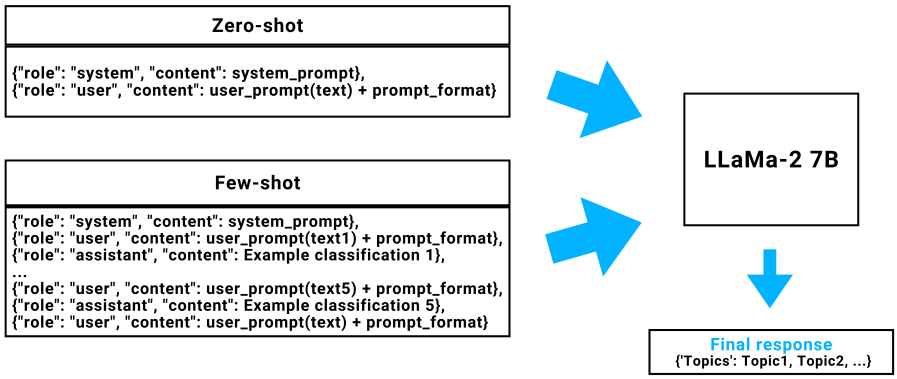

Figure 7: Zero-shot and Few-shot Prompting Strategies for Identifying Gold Labels

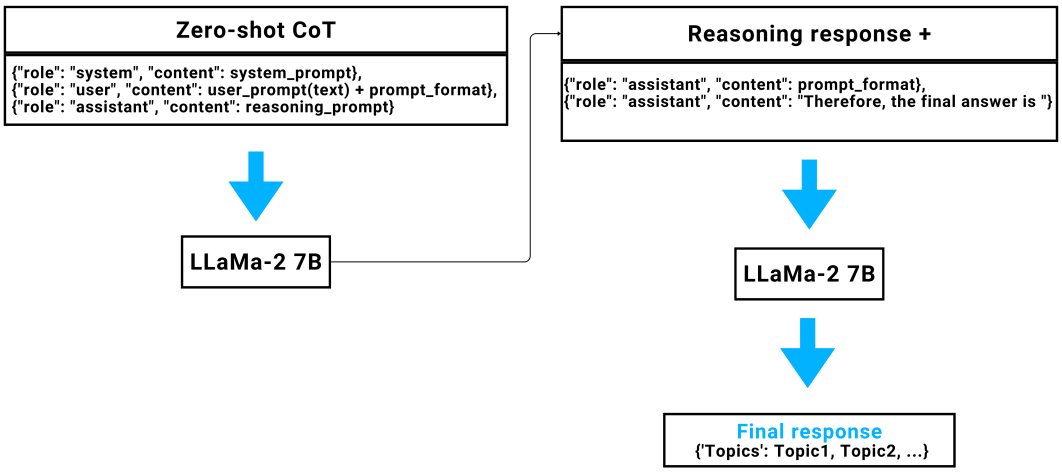

Figure 8: Zero-shot CoT Prompting Strategy for Identifying Gold Labels

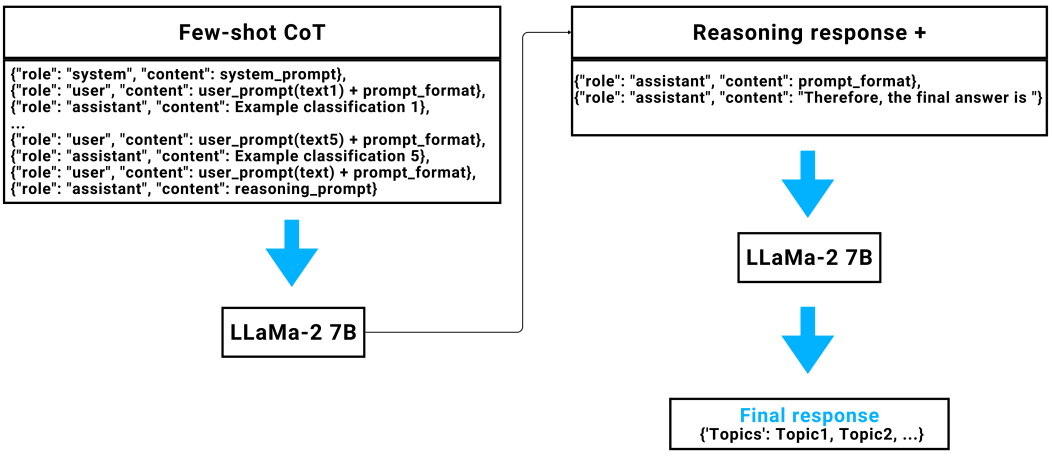

Figure 9: Few-shot CoT Prompting Strategy for Identifying Gold Labels

D Prompts structure for Exploring Human Label Variation

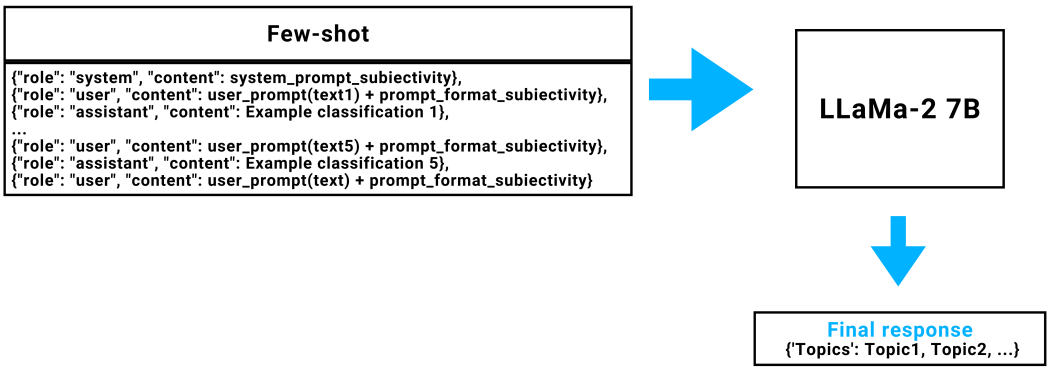

Figure 10: Few-shot Prompting Strategy for getting the labels per each annotator

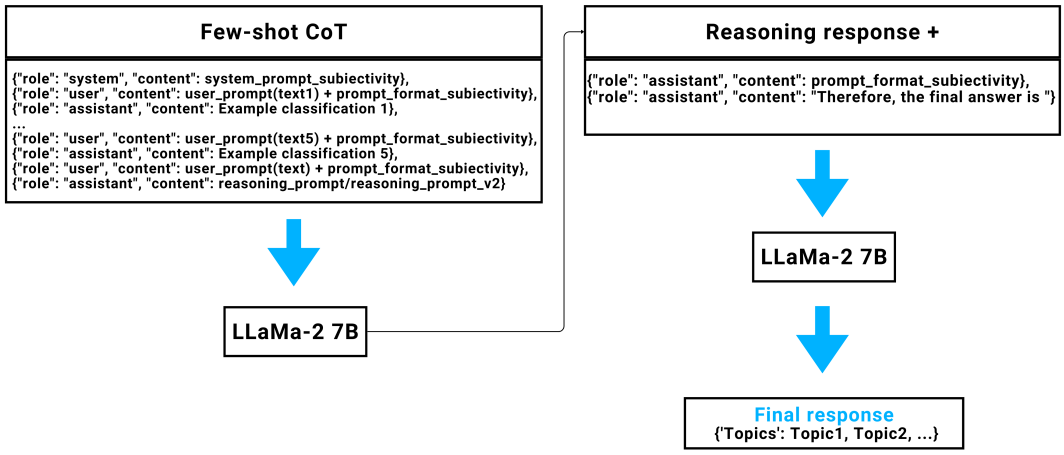

Figure 11: Few-shot CoT Prompting Strategy for getting the labels per each annotator

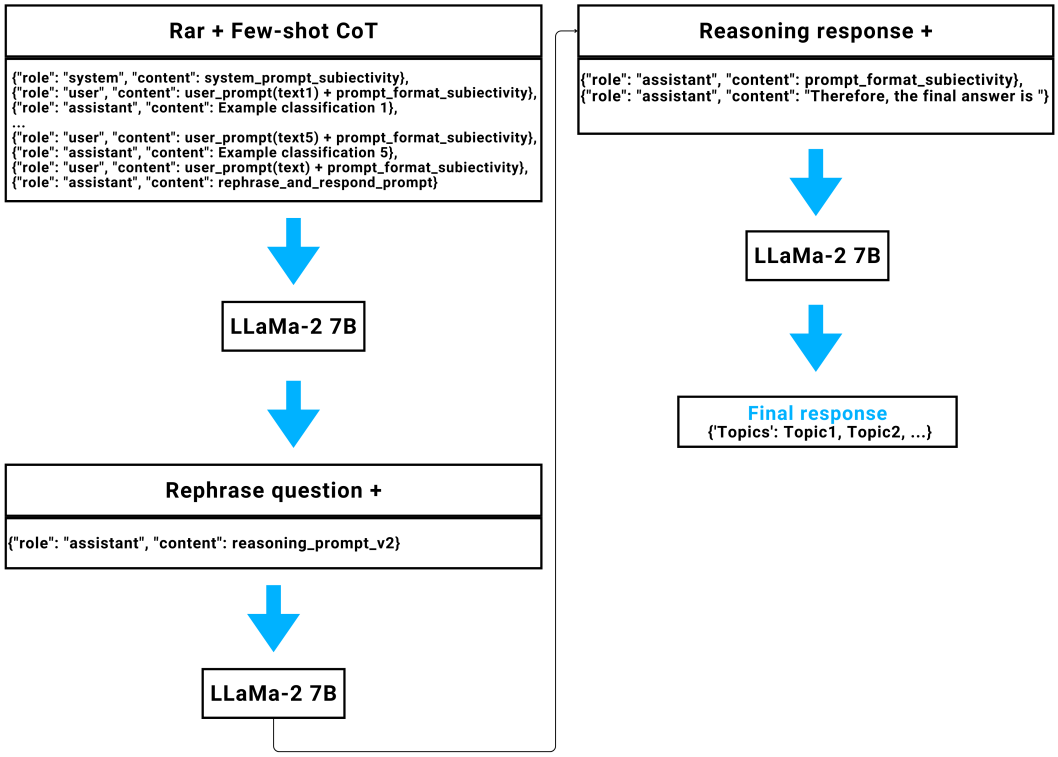

Figure 12: RaR + Few-shot CoT Prompting Strategy for getting the labels per each annotator

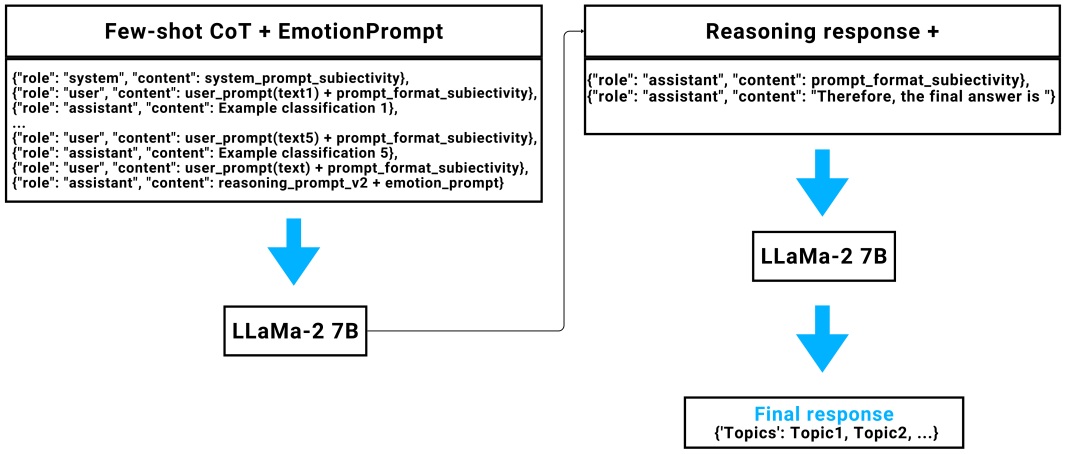

Figure 13: Few-shot CoT + EmotionPrompt Prompting Strategy for getting the labels per each annotator

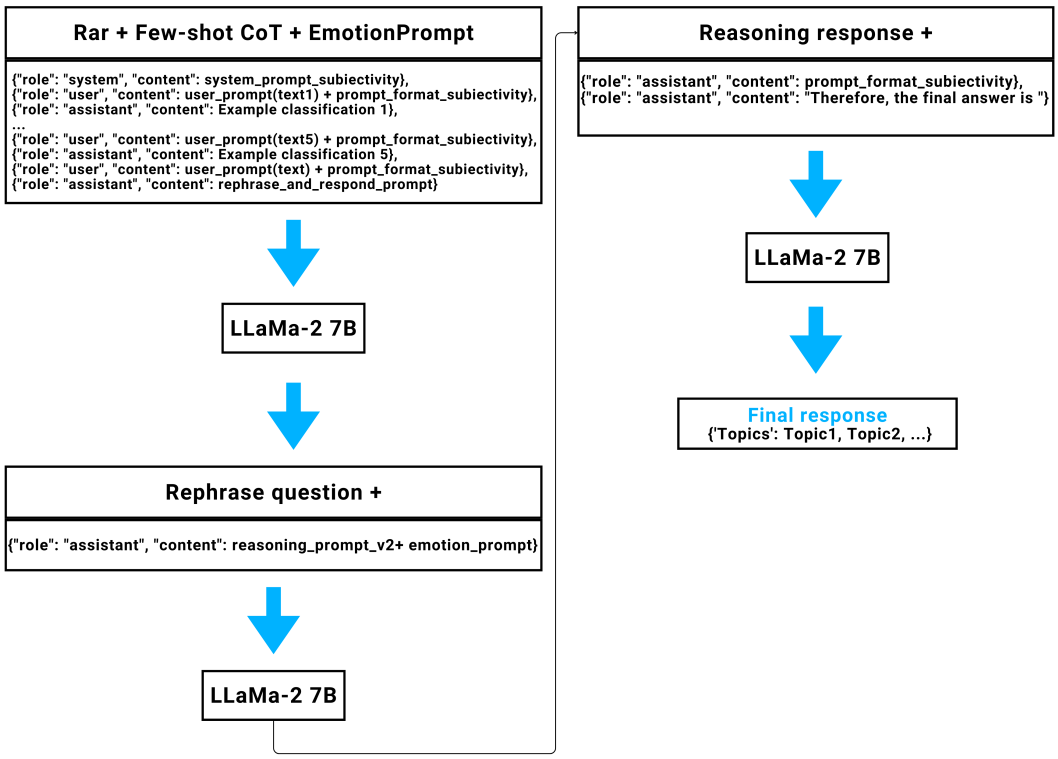

Figure 14: RaR + Few-shot CoT + EmotionPrompt Prompting Strategy for getting the labels per each annotator

E Statement on the use of generative AI

In writing this paper AI (ChatGPT) was used to improve the grammar, style, and/or spelling of the text. Moreover, AI assisted with LaTex formatting issues for tables. The prompt used are:

1. Check this text for grammar and spelling of the text.
2. Given table X, how can I format it so that it does Y in LaTex?