Using Large Language Models to Detect Deliberative Elements in Public Discourse

Detecting Subjective Emotions in Public Discourse

Bente Zuurbier
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: Bente Zuurbier
Final project course: CSE3000 Research Project
Thesis committee: Luciano Cavalcante Siebert, Amir Homayounirad, Enrico Liscio, Jie Yang

Abstract

In order to tackle topics such as climate change together with the population, public discourse should be scaled up. This discourse should be mediated as it makes it more likely that people understand each other and change their point of view. To help the mediator with this task, emotion detection can greatly help. Positive emotions can improve communications, while negative emotions cause people to be irrational and irritated. However, since emotions are highly subjective, it can make both predictions and evaluation more difficult.

Still, Large Language Models (LLMs) could be used to detect these subjective emotions using different prompting strategies and labels. The experiment included zero-, one-, fewshot and Chain of Thought (CoT) strategies. The precision was better for the one- and fewshot method compared to zeroshot. The CoT methods also showed an increase in precision, but a decrease in recall. The different labels were hard majority labels, soft labels and hard per annotator labels. In conclusion, providing examples improved the performance of the LLM. The CoT strategies were more precise, but gave a worse general prediction. The hard majority labels allow for more general predictions, where per annotator hard labels capture the perspective of different annotators. Soft labels reflect the subjective nature of the labels by providing probabilities instead of binary classification.

The experiment was done on a small data sample, so it is recommended to try the strategies on a larger data sample. Looking into appropriate evaluations for subjective predictions is also recommended in order to reflect the actual performance better.

1 Introduction

Public discourse allows people to come into contact with each other and exchange opinions. This is becoming increasingly important, as we face important choices regarding things like climate change or becoming more sustainable [1]. As you want the public to be involved with these choices, the discourse should take place on large scale. Some difficulties arise when doing so, as it can be difficult for a mediator to keep control of the exchange [2]. As such we should aim to make it easier for mediators to do their jobs well.

This research looks into how Large Language Models (LLMs) can be used to detect emotions in public discourse. Emotions play a large role in discourse, since they influence participants greatly. Negative emotions may cause participants to be distracted, manipulated and irrational[3]. The same research shows that positive emotions may help with understanding one another as well as communicating wants and needs. Therefore, emotion detection can greatly help mediators. If they are aware of the emotions of participants, they can use positive emotions to everyone's benefit and mitigate negative emotion before they do harm [4]. As LLMs have shown state of the art performance on natural language processing tasks, it is worth looking into their performance on emotion detection [5]. Since emotions are subjective, humans will likely annotate differently. It is especially interesting to see if the LLM is able to deal with that subjective nature of emotions.
From this follows the main research question:

How can Large Language Models be used to detect subjective emotions in public discourse?

The sub-questions are derived from the main question and are as follows:

- **RQ1**: How can a LLM be modelled to detect subjective emotions in public discourse?

- **RQ2**: What is the effect of different prompting strategies on the accuracy of subjective emotion detection in Dutch public discourse by a LLM?
- **RQ3**: What is the effect of different types of labels on the accuracy of subjective emotion detection in Dutch public discourse by a LLM?

These questions will be answered through multiple sections. First, the background information and related work sections take a look at what has already been done on this topic. **RQ1** is answered in the methodology. This section describes in detail how the annotation procedure was set up, how the LLM was prompted and how the experiment itself was done. After which the results of the experiment are presented, followed by the responsible research section and a discussion of the results. Lastly, the conclusions and main takeaways are summarized, followed by possible future work. After these sections, **RQ2** and **RQ3** are answered.

2 Background

In this section the most important related work is discussed as well as relevant definitions. First, public discourse and the role of emotions in it is discussed. After this, the emotion classification and subjective labels are explained.

2.1 Public Discourse

Public discourse can help people express their opinions and engage in politics. In effective discourse, people give reasons for their point of view and are actively involved. This is incredibly useful for increasing mutual understanding and making people change their point of view [6]. In order for public discourse to be effective, mediation is important. Otherwise we run the risk that people only talk to each other and do not listen to each other [1]. Such discussions generally do not help people understand each other better or change their point of view [6][1].

Public discourse can be found in many different places. It can take place physically, such as government debates, or it could take place online, on platforms like Reddit and X. Especially for such online platforms it is important that they are designed with deliberative elements, like emotions, in mind. Research shows that doing so increases equality and inclusiveness. It also allows citizens to engage more and express their opinions [2].

2.2 Emotions in Public Discourse

Detecting emotions can help a mediator know when to step in, either by diffusing negative emotions or using positive emotions to the groups benefit [3]. Research also showed that the proper handling of emotions can help build trust and the desire to work together again in the future. Too much of a focus on the other person's emotion can lead to excessive concessions however [7]. It is therefore important that the mediator is adept at recognizing and handling emotions, both their own and those of others. This is also called emotional intelligence.[3] [7]. As recognizing the emotions of yourself and others can be difficult to do, the help of a LLM has the potential to be useful in this process.

2.3 Classifying emotions

In order to detect and classify emotions with a LLM, there must first be a classification system which defines these different emotions. Paul Ekman was one of the firsts to argue that such a thing as "basic emotions" exist. Basic emotions are emotions everyone experiences, regardless of culture or societal background [8]. It started with a semantic scale of "pleasant" and "unpleasant". Later, they decided that a more nuanced scale of six distinct emotions would be better [9].

Recent research claims that many more distinct emotions exist. A study asked participants to map emotions to facial and body expressions[10]. They argued a minimum of 28 different categories is needed in order to capture the many nuances of emotions. It is important to be said that this study was done with English speaking participants from the US. This may mean that different cultures actually require more or less categories.

From these studies, the researchers of the GoEmotions dataset created an emotional taxonomy. It has 27 distinct emotions, with the addition of a "neutral" label [11]. This resulted in one of the largest human annotated datasets regarding emotion detection in text. Using the same taxonomy allows for easier potential comparison between the dataset used in this paper and the GoEmotions dataset. It also allows for easier comparison with past research. Many experiments have used the GoEmotions dataset to test their method of emotion classification.

2.4 Dealing with subjective labels

Something else to keep in mind with emotion detection is how to handle the subjectivity that is inherently present when annotating emotions. Emotions are a subjective feeling, so there often is no one true label [12]. Different methods on how to deal with absence of a ground truth have been researched, such as annotator and annotation embeddings [13]. The use of soft labels is also recommended. Soft labels assign a probability score to labels instead of a binary score [12]. When compared to using hard labels, the embeddings or soft labels generally performed better.

3 Related Work

With the emergence of LLMs, researchers started to look into the different uses for it. The studies focused on multiple areas.

Researchers started by comparing different existing models to LLMs. In one such study they compared GPT, a LLM, to IBM Watson, a system with the functionality to detect emotions. Both were asked to predict labels. Their conclusion was that the two models perform comparably [14]. It is important to note that GPT did have trouble with using the prescribed emotion classifications.

As it seemed that LLMs could predict labels reasonably well, studies started using LLMs to annotate data to train Machine Learning models. Often, properly annotated data is hard to come by, as annotating is expensive and time consuming. Using LLMs to create training data, they could train a new model or retrain an existing model within a specific domain [15]. A fewshot strategy was used to get the predictions. With this strategy the LLM is given a prompt, data to annotate and a small amount of annotated examples. This strategy can predict labels for an unannotated dataset. The dataset is then used to fine tune the model and improve performance, which is called pseudo-labelling[16]. This study showed using the pseudo-labels greatly improved performance compared to the baseline. Where the baseline model was only trained on the manually annotated examples.

Instead of using the predicted labels as training data, others looked at using the LLMs on their own. A study compared EmoBERTTiny, a non-generative LLM fine-tuned to detect emotions, to Llama 2 and Mistral, which are generative LLMs. Their results showed that EmoBERTTiny outperforms Llama 2 and Mistral considerably. The main focus of their research was to find a LLM that could detect emotions fast and accurately as they wanted it the be usable in real time [17]. It should be noted that they only used zeroshot, oneshot and three-shot prompting for Llama and Mistral, so perhaps other methods, such as Chain of Thought reasoning, would perform better.

The next step was to look at the power of combining LLMs. A study tried combining generative AI models, such as ChatGPT, and fine tuned domain specific models, such as RoBERTa. The aim being to create a model that can detect emotions better than any single LLM can [18]. It resulted in a model that could detect broader emotional context than a specifically trained model can. This model had accurate results and lower training costs.

All in all, different studies have been performed on the usage of LLMs to detect emotions. They have focused on comparing LLMs to pre-trained models, using LLMs to create more annotated data and combining LLMs. There is not a lot of research into the impact of different prompting strategies yet, beyond the comparison of zeroshot and fewshot. This is looked into more in this paper, which compares the performance of zero, one, few-shot and Chain of Thought strategies. Furthermore, the effect of different kinds of labels on the accuracy of emotion detection using LLMs has not been studied thoroughly yet. As such, this paper also looks at majority hard labels, soft labels and per annotator labels.

4 Methodology

This section describes how the data was annotated and how the obtained labels are aggregated. It also outlines the prompting strategies, the set-up of the experiment and the chosen evaluation methods. Figure 1 shows an overview of this all.

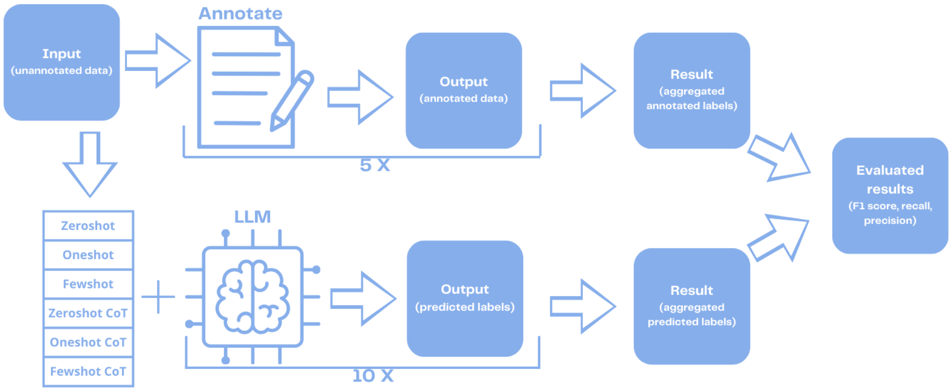

Figure 1: Overview methodology

4.1 Annotating the Data

The data was from a survey done in South West Friesland about the future energy policy. This research was done by TU Delft and more information can be found here: Friesland Study. The respondents were Frisian and this experiment was run not on the original text, but on the translated text. The dataset for this research was not yet annotated. So, the classification system and annotation procedure had to be picked.

The emotion taxonomy of GoEmotions [11] was given to the annotators. These 28 labels, as well as English text to be annotated, were given to the annotator in an Excel sheet. Annotators were instructed to read the sentence, multiple times if needed, and type a 1 in the column of the emotion they believed could be detected in the sentence.
In order to better deal with annotator bias, multiple people annotated the data. In this case five students of different nationalities annotated the data. For none of the students the English language was their first language, but all spoke it fluently.

4.2 Aggregating the Labels

Once the data was annotated, the next step was aggregating the labels. With objective labelling tasks, using majority vote is often a good choice. Most of the time, the majority will have made the right choice. However research into the topic of subjective labelling tasks suggests using soft labels instead [12].

Different labels were used in the experiments. First a hard multi-label approach was used, where the labels are aggregated. If at least two annotators chose the label, it was considered "correct". Another use of hard multi-labels was by defining all the annotated labels to be correct. The second approach used soft labels, meaning the labels were given a probability for being correct. All annotated labels were added together and divided by the amount of annotators to create the probabilities. The last approach was a hard multi-label per annotator. Instead of aggregating the labels, this approach tries to capture the subjectivity per annotator.

4.3 Prompting Strategies

In order to detect emotions using a LLM, prompting strategies had to be chosen.
The zeroshot, oneshot and fewshot strategies were chosen as they can be used with relatively little data. All provide the same prompt and data to the LLM. With oneshot and fewshot the LLM is also provided with examples. This helps to give it some context and provides a very small training sample to base the answer on. The examples were chosen based on the diversity of the labels represented by them. These strategies were used to predict both hard and soft labels.

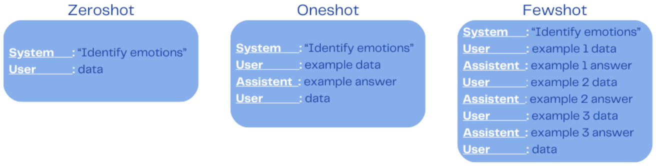
Zeroshot
System : "Identify emotions"
User : data
Oneshot
System : "Identify emotions"
User : example data
Assistent : example answer
User : data
Fewshot
System : "Identify emotions"
User : example 1 data
Assistent : example 1 answer
User : example 2 data
Assistent : example 2 answer
User : example 3 data
Assistent : example 3 answer
User : data

Figure 2: Zeroshot, oneshot and fewshot

Furthermore, the Chain of Thought (CoT) prompting strategies are used. CoT asks the LLM to think about the steps it takes to reach the answer. The first call asks for the reasoning behind the answer and the second call asks to extract the labels from this reasoning. The differences between zero-, one-, and fewshot can be seen in figure 3. The benefit of CoT is that by asking the LLM to explain its steps, it is able to do complex tasks better.

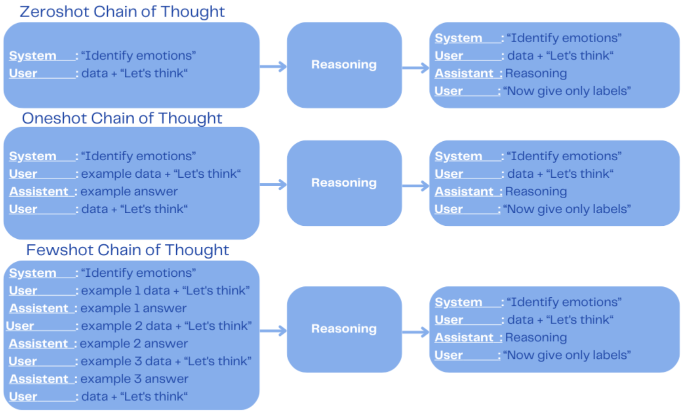
Zeroshot Chain of Thought
System : "Identify emotions"
User : data + "Let's think"
Reasoning
System : "Identify emotions"
User : data + "Let's think"
Assistant : Reasoning
User : "Now give only labels"
Oneshot Chain of Thought
System : "Identify emotions"
User : example data + "Let's think"
Assistent : example answer
User : data + "Let's think"
Reasoning
System : "Identify emotions"
User : data + "Let's think"
Assistant : Reasoning
User : "Now give only labels"
Fewshot Chain of Thought
System : "Identify emotions"
User : example 1 data + "Let's think"
Assistent : example 1 answer
User : example 2 data + "Let's think"
Assistent : example 2 answer
User : example 3 data + "Let's think"
Assistent : example 3 answer
User : data + "Let's think"
Reasoning
System : "Identify emotions"
User : data + "Let's think"
Assistant : Reasoning
User : "Now give only labels"

Figure 3: Zeroshot, oneshot and fewshot Chain of Thought

The final strategy that was looked at is fewshot per annotator. Instead of general examples, the LLM is provided with three examples from each annotator and asked to predict labels per annotator.

An overview of the different strategies can be seen in table 1.

Method	**Examples Given**	**Reasoning Step**	**Per Annotator Prediction**
Zeroshot	0	No	No
Oneshot	1	No	No
Fewshot	3	No	No
Zeroshot CoT	0	Yes	No
Oneshot CoT	1	Yes	No
Fewshot CoT	3	Yes	No
Fewshot per annotator	3	No	Yes

Table 1: Overview of the prompting strategies

There are multiple elements to the prompts of these strategies. There are different roles, namely system, user and assistant, that you can provide. Along with these roles, content is provided. In the prompt, the system role gives the task and user role provides the text to be labelled [19]. The same paper proposes to use the assistant role when providing an example answer, which both oneshot and fewshot use.
The content part of each strategy was created by using a prompt similar to the one in the paper and asking the LLM itself to improve it. It is also asked to answer in JSON format as it creates more uniform results. This simplifies the processing of the predicted labels.

Specifically for the CoT strategies, a reasoning response was needed for the content of the assistant role. This content was created by asking the LLM for the reasoning multiple times. A single one of these responses was chosen, which was selected on the structure, number of steps and if it generally made sense or not. The chosen response had three steps, asking the LLM to consider the tone, phrases that convey emotions, and overall sentiment. These steps look both at the overall tone and sentiment of the text, as well as words that could be connected to specific emotions.

4.4 Experimental Set-up

For this experiment, it was important to run the LLM locally, as the data used is not public. Ollama was chosen, as it is free to use and offers multiple open source LLM models. The Llama3 model was chosen as it works well with the different prompting strategies described below. It is also the most recent opensource model from Meta.

With the created prompts and annotated data, the experiment was conducted. Using Python, code was created to prompt the LLM, process the response and evaluate the results.

The annotated data was aggregated into a numpy array. This format is needed for the f1, recall and precision methods of the sklearn library. The next step was to create a dataframe of the unannotated data. The LLM is run on this dataframe, using the different prompting strategies. As an LLM returns different results each time, it was run ten times. The results were then aggregated for more robust results.
Once both the annotated labels and predicted labels arrays are created, the results can be evaluated. This is done by calculating the micro F1 score, recall and precision.

4.5 Evaluation Metrics

Several evaluation metrics were chosen in order to create a good reflection of the results.

- **Precision**: this tells us how many of the predicted labels are actually correct. This is done by dividing the true positives by both the true positives and the false positives.
- **Recall**: shows how many of the positive labels are actually found by the LLM. This is done by dividing the true positives by both the true positives and false negatives.
- **F1 score**: gives an overall score for both precision and recall. Specifically the micro F1 score is used, as this works better for multi-label problems.
- **Fleiss Kappa**: measures the level of agreement between the annotators. This score looks at the agreement between annotators compared to the level of agreement they would get by pure chance.

5 Results

The results are organized per labels. First, the results using hard majority labels are discussed, followed by the soft probabilistic labels and ending with the subjective per annotator labels.

5.1 Results of Hard Majority Labels Aggregated

The following results are for the zeroshot, oneshot, fewshot and CoT training methods. The annotated labels are considered if at least two annotators picked it. For the predicted labels, the results are aggregated, and if at least two runs predicted the label, it is considered.

Method	Micro F1 Score	Micro recall	Micro Precision
Zeroshot	0,385	0,420	0,355
Oneshot	0,469	0,580	0,394
Fewshot	0,486	0,537	0,444
Zeroshot CoT	0,410	0,399	0,422
Oneshot CoT	0,495	0,558	0,445
Fewshot CoT	0,480	0,485	0,474

Table 2: F1 score, recall and precision for all training methods

As can be seen in table 2, the oneshot and fewshot methods perform better than zeroshot in all the metrics. Providing the LLM with examples allows for better overall recall and precision. Furthermore, the CoT methods perform better than their other counterparts on precision. This is likely due to the fact that it generally predicts less labels.

5.2 Results of Hard Labels Per Run

Other results include the amount of correct and incorrect labels. The predicted label is considered correct if at least one annotator chose that label. These labels are not aggregated, but considered per run. These metrics show whether a LLM can predict labels that humans would possibly annotate.

A total of 329 unique labels were given by the annotators. The agreement between the annotators was low, as indicated by a Fleiss Kappa score of 0.00365. Even though the agreement was low, the choices of the annotators should still be considered as the truth. For subjective annotating such as emotion annotation, there is no objective "right" or "wrong". If you discard the labels that were not chosen often, you would be undermining the contributions of those annotators [13].

	Precision		**Correct Labels**		**Incorrect Labels**	
	M	SD	M	SD	M	SD
Zeroshot	0,660	0,0248	54,4	3,720	28	2,145
Oneshot	0,720	0,0147	93	2,145	36,1	2
Fewshot	0,768	0,0304	75,8	2,857	23	3
Zeroshot CoT	0,709	0,0441	39,5	3,722	16,3	3,132
Oneshot CoT	0,718	0,0221	68,5	2,377	27,0	2,864
Fewshot CoT	0,764	0,0244	52,6	3,137	16,3	2,492

Table 3: Precision, correct and incorrect labels for all strategies with mean (M) and standard deviation (SD)

Table 3 shows the precision of the different strategies. Oneshot and fewshot have a higher precision than zeroshot. The precision of their CoT counterparts is very comparable. The main difference between them is that CoT predicts less labels in total.
The precision scores of the individual runs can be found in the appendix 6. Zeroshot has only one run which performs better than the worst oneshot run. The same is true for oneshot

when compared to fewshot. Zeroshot CoT has a wide spread of precision scores, likely due to the differences in reasoning provided by the LLM without an example.

5.3 Results of Soft Probabilistic Labels

The results below are for probabilistic soft labels. The LLM was asked to provide a probability score for every predicted label. The annotated labels were added together and divided by five to create probability score for the annotated labels.

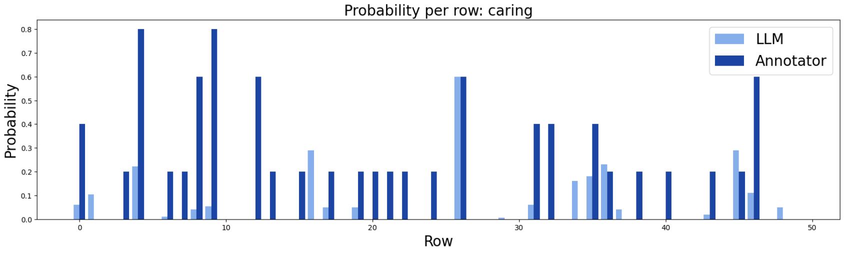

Figure 4: Soft labels for the caring label predicted by zeroshot

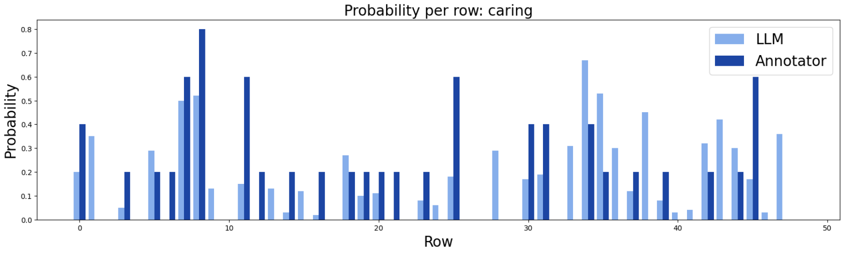

Figure 5: Soft labels for the caring label predicted by oneshot

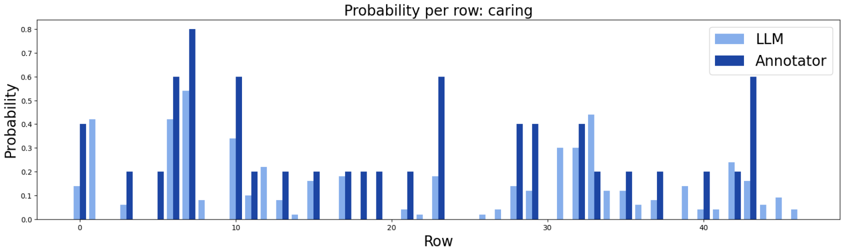

Figure 6: Soft labels for the caring label predicted by fewshot

Figures 4, 5 and 6 show the probability for the caring label predicted by a zeroshot, oneshot and fewshot method. As caring is a label often given by annotators, the LLM had more training material each run. So for this specific label, it can be observed that the scores become more accurate when more examples are given. It is not the case for all labels however, as not enough training data was available to train the model extensively.

5.4 Results of Subjective per Annotator Labels

The results below are from asking the LLM to predict labels per annotator specifically.

Annotator	Annotated	Predicted	Labels Given to LLM
Annotator 1	81	161	8
Annotator 2	115	188	6
Annotator 3	96	219	7
Annotator 4	95	233	8
Annotator 5	143	213	11

Table 4: Number of annotated and predicted labels per annotator

In table 4, a label is considered predicted if at least two runs of the LLM predicted that label. The same examples were used as for the other fewshot methods. As such, it differed on how many labels were given in total to the LLM to learn the annotator perspective. However, this does not seem to have a clear connection to the amount of labels predicted.

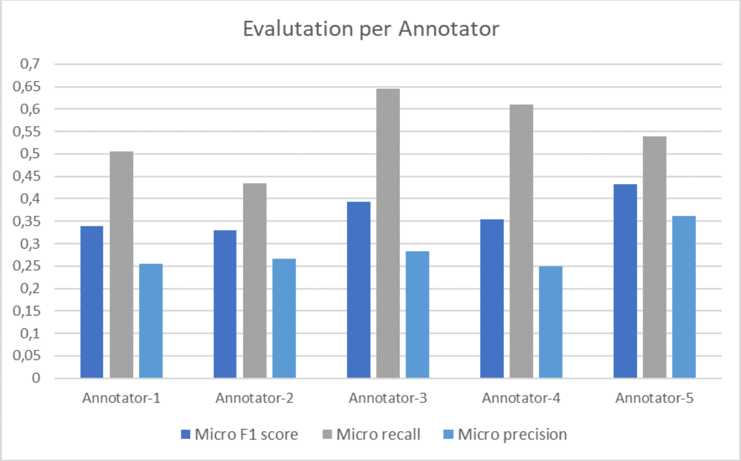

Figure 7: F1 score, recall and precision for per annotator training

As the amount of labels predicted is much higher than the amount of labels annotated, for most of the annotators the recall is quite good. Figure 7 also shows that this comes at the expense of the precision, as only about a quarter of the predicted labels is actually correct.

	F1 score		**Recall**		**Precision**		**Labels**	
	M	SD	M	SD	M	SD	M	SD
Annotator 1	0,327	0,0294	0,372	0,0376	0,293	0,0246	102,7	3,716
Annotator 2	0,278	0,0388	0,247	0,0353	0,318	0,0444	89,3	3,848
Annotator 3	0,313	0,0200	0,316	0,0224	0,310	0,0198	97,8	4,308
Annotator 4	0,269	0,0223	0,287	0,0258	0,253	0,0219	107,9	5,890
Annotator 5	0,342	0,0271	0,287	0,0253	0,423	0,0301	97,0	4,171

Table 5: F1 score, recall, precision and number of labels per annotator with mean(M) and standard deviation(SD)

Another way of looking at the results is by looking at the averages per run, instead of aggregating the results. The evaluations of that can be seen in table 5. Both the average recall and precision are low, which is consistent over all the runs. In this case, aggregating the results leads to higher scores.

6 Responsible Research

As it is important to conduct your research responsibly, factors such as reproducibility and integrity were taken into account when conducting the research for this paper.

In order to ensure better reproducibility, several steps were taken. The code is publicly available at the following Github repository: Github Code. The code is provided with comments and documentation in order to understand how the code works. Furthermore, the experiments were run 10 times to create more reproducible results. As a LLM gives different answers when it is run, the exact results are not reproducible. The general trends were stable however. You could possibly set the temperature to 0 or give a specific seed in order to always get the same results. That however means a lot of the creativity is lost. This would not result in results reflective of the actual capabilities of the LLM and was therefor not chosen. These steps try to uphold multiple principles of the Netherlands Code of Conduct, namely Honesty, Scrupulousness and Transparancy.

It is however important to note that the data used in this experiment is not publicly available. This dataset was used as it is public discourse about sustainability. As well as that each student researching different deliberative elements could use the same data. The data itself was properly anonymized to protect the identity of participants. Only ID's of participants are available. Further information and contact information can be found here Friesland Study. As the LLM is barely trained on the examples itself, it can however still be said that this experiment could be done on a different dataset with similar results. By following the annotation procedure as described and choosing the examples for oneshot and fewshot as described, similar results should be possible.

Finally, this research was not done alone. It was done under guidance of the supervisors and with sparring with other students. The research could have been influenced by this, but no scientific integrity was sacrificed because of it. The research was also conducted in order to graduate, but this does not make it any less socially relevant. If LLMs can be used to help mediate discourse, it can have a positive impact on society. It could help with large scale debate on important topics, such as climate change and sustainability. As such the principle of Responsibility was also taken into account.

7 Discussion

The constraints will be discussed first. As the data was annotated by students, there were not many data points that could be used. The data was also translated from Dutch, which could mean there were translation errors. This possibly makes it more difficult for the LLM to predict emotions. This is however an accurate depiction of real world application, as not all sentences in public discourse are correct English.

The results show that providing the LLM with examples allows for more accurate predictions.

Both the precision and amount of correct labels of oneshot and fewshot is higher than that of zeroshot. Fewshot has the highest precision, so providing more examples makes the LLM more precise. The standard deviation over the ten runs is low, so the results are statistically valid.
For recall however, oneshot performs best. The format of given answers is more uniform because an example was given. Yet a larger amount of unique labels is still given. As such, it is more likely that more annotated labels are predicted.

In case of CoT, the precision of zeroshot CoT is better than zeroshot. This could indicate that reasoning does benefit the precision. The difference is less significant between oneshot, fewshot and their CoT counterparts. Either the number of examples has a higher impact on precision or the provided reasoning examples for the CoT methods should have been different.

Moreover, the LLM seems capable of assigning soft labels. It is difficult to say they perform better, as the F1 score, recall and precision cannot be used for soft labels. Perhaps with more extensive training, the predicted labels can become more accurate.

Furthermore, it is difficult to say if an LLM can predict labels from a specific annotator's perspective using fewshot. Aggregating the results does seem to help, but in order to truly capture an annotator perspective more training data is needed. Providing the LLM with more examples could possibly help capture the subjective nature of emotion annotation from different annotators better.

It can also be argued that using objective evaluation methods for a subjective task does not give the true picture. As the annotators had little agreement between them, it cannot be said a true label exists. A predicted label is incorrect if no annotator chose it, yet the label often makes sense when compared to the text.

This is illustrated by the following example:

"Residents are needed. It's also a lot of fun to get involved. However, there are quick choices of principle that exclude other solutions. Very valuable on a small scale (including mienskip) and that certainly in combination with climate and reuse. Local initiatives can also lead to fragmentation while electricity must always be available, and inefficiencies."

- **Annotator 1:** amusement, confusion, optimism
- **Annotator 2:** approval, excitement, realization
- **LLM prediction:** annoyance, approval, caring, confusion, disapproval, neutral

Both of the annotators give different labels, yet you would not say either of the humans is wrong. The LLM predicted some of the annotated labels and also some there were not chosen. Those labels however also make sense. As such, it can be argued that using objective evaluation metrics for a highly subjective task does not reflect the true capabilities of the LLM.

8 Conclusions and Future Work

Both the takeaways and future work are described in this section.

How an LLM can be modelled to detect emotions, **RQ1**, is comprised of choosing the annotation process, an existing model and prompting strategy. For annotation, a common classification system and instructions are needed. The chosen model was Llama3, as it is open-source and one of the most recent models. It also worked well with the different training methods and gave consistent results. The prompting strategies chosen were zeroshot, oneshot, fewshot and Chain of Thought. Another strategy was fewshot per annotator, where the LLM predicted labels per annotator.

For **RQ2**, the effect of prompting strategies on the accuracy of the model had the following conclusions. Providing the LLM with examples improved the precision and recall of the predictions, making fewshot the better method. Chain of Thought methods could be better for unannotated data, as zeroshot CoT had the largest increase in performance. Lastly, using only fewshot, the LLM could not properly predicting labels from an annotator's perspective. While aggregating the data did help the performance, it is likely that more training data is needed in order to properly capture an annotator's perspective.

For **RQ3**, it can be concluded that hard labels allowed the use of evaluation metrics such as F1 score, recall and precision. The majority vote labels allowed for general predictions, where the individual hard labels allow to take the subjectivity and bias per annotator into account. The soft labels allow for more information to be given to the LLM to train on. It also reflects the subjective nature of the task better.

As such, the answer to how LLMs can be used to detect subjective emotions in public discourse is through different prompting strategies and labels. A LLM can be used to help detect emotions, but it is important to keep in mind how subjective the emotions are. As the very nature of subjectivity is that there are no right or wrong answers, the predictions should not be taken as the truth. The LLM is as much "right" as a human annotator is. As long as that is taken into account, the LLM can be used to help with emotion detection.

Potential future work is to run the code on the GoEmotion dataset. As discussed, the data was annotated by students and was translated from Dutch. Perhaps running the code on GoEmotion leads to different results. Another constraint was time, as such there are other training methods that can still be tried, such as finetuning. Furthermore, the performance of different LLMs than Llama3 could be look at.

Other potential research is looking at how we should evaluate the performance of an LLM on a subjective task. Perhaps a user study could be done, where they are given two sets of emotion labels for a piece of text. One set annotated by a human and the other set predicted by a LLM. Perhaps humans prefer one set over the other, or perhaps LLMs can annotate on the same level as the average human can.

References

[1] J. Forester, “Challenges of deliberation and participation”, *Les ateliers de l’éthique*, vol. 1, no. 2, pp. 19–25, 2018. DOI: https://doi.org/10.7202/1044678ar.

[2] R. Shortall, A. Itten, M. Meer, P. Murukannaiah, and C. Jonker, “Reason against the machine? future directions for mass online deliberation”, *Frontiers in Political Science*, vol. 4, 2022. DOI: https://doi.org/10.3389/fpos.2022.946589.

[3] E. Kelly and N. Kaminskienė, “Importance of emotional intelligence in negotiation and mediation”, *International Comparative Jurisprudence*, vol. 2, no. 1, pp. 55–60, 2016. DOI: 10.1016/j.icj.2016.07.001.

[4] N. Yeend, “Creating an environment where participants can express their emotions in constructive ways”, pp. 1–3, 2021. [Online]. Available: https://plaintiffmagazine.com/recent-issues/item/emotions-in-mediation.

[5] R. Venkatakrishnan, G. M., and M. Canbaz, “Exploring large language models’ emotion detection abilities: Use cases from the middle east”, *2023 IEEE Conference on Artificial Intelligence (CAI)*, pp. 241–244, 2023. DOI: 10.1109/CAI54212.2023.00110.

[6] E. Schneiderhan and K. Schamus, “Reasons and inclusion: The foundation of deliberation”, *Sociological Theory*, vol. 26, no. 1, pp. 1–24, 2008. DOI: 10.1111/j.1467-9558.2008.00316.x.

[7] K. Kim, N. Cundiff, and S. Choi, “The influence of emotional intelligence on negotiation outcomes and the mediating effect of rapport: A structural equation modeling approach”, *Negotiation Journal*, vol. 30, no. 1, pp. 49–68, 2014. DOI: 10.1111/nejo.12045.

[8] P. Ekman, “Are there basic emotions?”, *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992. DOI: https://psycnet.apa.org/doi/10.1037/0033-295X.99.3.550.

[9] P. Ekman, “An argument for basic emotions”, *Cognition and Emotion*, vol. 6, no. 3/4, pp. 169–200, 1992. DOI: https://doi.org/10.1080/02699939208411068.

[10] A. S. Cowen and D. Keltner, “What the face displays: Mapping 28 emotions conveyed by naturalistic expression”, *The American psychologist*, vol. 75, no. 3, pp. 349–364, 2020. DOI: 10.1037/amp0000488.

[11] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and R. Sujith, “Goemotions: A dataset of fine-grained emotions”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020. DOI: 10.18653/v1/2020.acl-main.372.

[12] 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. DOI: https://doi.org/10.1613/jair.1.12752.

[13] N. Deng, X. Zhang, S. Lui, W. Wu, L. Wang, and R. Mihalcea, “You are what you annotate: Towards better models through annotator representations”, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12 475–12 498, 2023. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.832.

[14] D. Carneros-Prado, L. Villa, E. Johnson, C. Dobrescu, and A. Barragán, “Comparative study of large language models as emotion and sentiment analysis systems: A case-specific analysis of gpt vs. ibm watson”, *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023)*, pp. 229–239, 2023. DOI: 10.1007/978-3-031-48642-5_22.

[15] G. Tu, B. Liang, B. Qin, *et al.*, “An empirical study on multiple knowledge from chatgpt for emotion recognition in conversations”, *Findings of the Association for*

Computational Linguistics: EMNLP 2023, pp. 12 160–12 173, 2023. DOI: 10.18653/v1/2023.findings-emnlp.813.

[16] U. Malik, S. Bernard, A. Pauchet, C. Chatelain, and J. Cortinovis, "Pseudo-labeling with large language models for multi-label emotion classification of french tweets", *IEEE Access*, pp. 15 902–15 916, 2024. DOI: 10.1109/ACCESS.2024.3354705.

[17] W. Stigall, M. Al Hafiz Khan, D. Attota, F. Nweke, and Y. Pei, "Large language models performance comparison of emotion and sentiment classification", *Proceedings of the 2024 ACM Southeast Conference on ZZZ*, pp. 60–68, 2024. DOI: 10.1145/3603287.3651183.

[18] B. V. Kok-Shun, J. Chan, G. Peko, and D. Sundaram, "Intertwining two artificial minds: Chaining gpt and roberta for emotion detection", *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pp. 1–6, 2023. DOI: 10.1109/CSDE59766.2023.10487718.

[19] 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", 2023. DOI: https://doi.org/10.48550/arXiv.2401.00284.

A Appendix

A.1 Results hard labels

Zeroshot	0,622	0,63	0,646	0,65	0,652	0,654	0,671	0,68	0,694	0,701
Oneshot	0,695	0,699	0,706	0,72	0,722	0,722	0,732	0,733	0,736	0,738
Fewshot	0,699	0,75	0,752	0,758	0,765	0,765	0,783	0,79	0,802	0,812
Zeroshot CoT	0,608	0,667	0,698	0,698	0,707	0,72	0,722	0,745	0,759	0,765
Oneshot CoT	0,684	0,686	0,691	0,717	0,718	0,723	0,731	0,734	0,745	0,747
Fewshot CoT	0,725	0,732	0,746	0,754	0,754	0,771	0,779	0,789	0,794	0,797

Table 6: Precision scores of all prompting strategies