Proposal to make use of prompting of LLMs to get better classification results than Step 3:

We will try to use the strong generalization capabilities of an LLM like GPT-4 to try to directly classify documents from the datasets.

After a few trials using ChatGPT, it became clear that GPT-4 has great potential at zero-shot classification. If given a prompt to classify amongst given options & the first few documents it correctly classifies them. However if we say "do the same task as before on the following documents" in subsequent prompts occasionally it gives out different classes than the given options. One way is to reinforce the options by 'reminding' GPT-4 of the available options. Or, by using the same original prompts for each document (via API call on each document).

The prompt used for each document is:

News:

```
"Categorize this document from a news article dataset as either one of
these classes - *Business* or *Technology* or *Sports* or *Politics*.
Choose one (the most likely) of the 4 labels I provided the
document/review. Pay particular attention to the core idea in the document
for deciding which class it belongs to. Return only the class name, i.e.
*Business* or *Technology* or *Sports* or *Politics*. Also note that
everything not separated by a newline character is one news
article/document: <document>"
```

Movies:

```
"Categorize this document from a movie review dataset as either one of -
*bad* or *good*. Choose one (the most likely) of the 2 labels I provided
the document/review. Pay particular attention to the core idea in the
document for deciding which class it belongs to. Return only the class
name, i.e. good or bad. Also note that everything not separated by a
newline character is one review/document: <document>"
```

	Step 3 -- F1 score	**Step 4 -- F1 score**
News dataset (validation)	**0.82** (with CatE embeddings)	**0.88**
Movies dataset (validation)	**0.75** (with W2V embeddings)	**0.94**

We see that GPT-4 is able to do way better than Step 3 on the movie dataset, while the delta/change isn't much with the news dataset (where options & thus perplexity could be more).

We see an increase in this performance could be attributed to GPT-4's strong generalization capabilities and an effect prompt.