Thematic classification of text using LLMs

Jason Carvalho

February 5, 2024

1 Text Classification

1.1 Introduction

This small study, undertaken as part of the wider CHILD pilot, focuses on harnessing LLM technology to classify existing text extracts within LED, a task traditionally performed by human domain experts, to address the challenges posed by the volume of textual data in fields such as music history.

Our experiment evaluates the effectiveness of an LLM in categorizing text extracts under the specific theme of childhood, comparing its performance with that of a human domain expert. The comparison aims to quantify the alignment between machine and human interpretations in textual analysis, look at areas where LLM technology may show weaknesses and also investigate if there areas where LLMs are able to shed new light on data that may go unnoticed by humans.

1.2 Methodology

To begin this experiment, we took a sample of around 1000 data items from the LED dataset that had already been reviewed manually by a domain expert; a music historian at The Open University. A number of these had been marked by the expert as describing the listening experiences of children, describing childhood or falling under the theme of childhood in some way. Our plan to was develop a suitable prompt to ask an LLM if it considered each of the listening experiences to fall under the thematic banner of childhood.

To perform these experiments, we used OpenAI's API access to their ChatGPT-3.5 model, making one call per entry from an iterative Python script. After some manual testing on a selection of LED entries, the following LLM prompt was agreed upon:

> 'Does the following passage cover the theme of childhood/youth or describe or mention childhood/youth or children and young people in any way?'

The LLM was asked to not only provide a true/false answer, but also to explain its reasoning. It was also asked to provide its answers and reasons in a JSON structure adhering to a pre-defined schema such that processing and rendering of the results could be automated.

An example response given by the LLM:

```
{
    Childhood: True
    Reason: "The passage describes a little boy's excitement
        and delight while watching a regiment band rehearsal.
        The mention of the little boy, his interaction with
        his father, and his enthusiasm for the music
        indicates a focus on childhood/youth."
}
```

Tuning parameters were largely left at default values and no specific configurations were used.

1.3 Results

The initial results from this experiment (Table 1) were somewhat problematic. Not only did the LLM miss a significant number of the entries flagged by the domain expert, it also flagged a large number of its own that were not originally marked by our expert.

Some probing into what appeared to be false positive results revealed that many of these entries had in fact been correctly identified by the LLM, highlighting the difficulties faced by humans in exhaustively categorising large volumes of textual data. The LLM had also missed a number of entries too, so a slightly different approach to the LLM prompt was taken. As part the wider CHILD project, a scenario was devised that describes a typical use case for the pilot.

The Scenario:

> ‘Ortenz wants to characterize children’s experience of music as witnessed in bibliographic and artistic sources. She is looking for primary sources (e.g. Personal journals, literary texts) wherein to find evidence of listening experiences. She needs to collect and analyze large corpora of texts and images recording or depicting children’s experience with music. Documents include official sources (e.g. newspaper articles, reviews of concerts, paintings) and sources produced by “ordinary people”. She prefers the latter as they provide more reliable feedback, and she looks at the context of production of such sources (where, when, who created the source, the goal, which related events exist), contents (recurring motifs and themes), and elicited emotional responses. She collects sources belonging to different historical periods so as to characterize the development of identified phenomena.’

For our second attempt at thematic classification using an LLM, we instead presented the model with the above scenario and asked the LLM if each passage of text helps to address and answer the requirements of the scenario. This experiment yielded more positive results than the previous prompt, with some overlap but also with many more passages now identified as falling under the banner of childhood.

Combining the results of the original prompt and also the second scenario-based prompt gave a total of 85 LED entries that had been classified as childhood-related by the LLM.

Table 1 shows the results of the thematic classification. The initial prompt results are indicated as (1) and the second classification with the updated scenario-based prompt is shown as (2).

Total listening experiences analysed	878
Flagged by domain expert	26
Flagged by ChatGPT-3.5 (1)	45
Flagged by ChatGPT-3.5 (2)	71
Flagged by ChatGPT-3.5 (total)	85

Table 1: Classification results

Since the original set of domain expert-identified entries showed to be incomplete, we employed the same music historian again to perform a qualitative evaluation the results of the LLM classification. They were presented with around 100 entries, specifically chosen where the LLM had flagged entries that were previously not in the original expert-selected subset. The purpose of the evaluation was to investigate the reasons for the discrepancies, rather than measure precise quantitative accuracy.

Our expert was asked to review each entry and state whether they agreed or disagreed with the LLM’s inclusion of that entry in the CHILD subset. Where they disagreed, they were asked to explain their reasoning.

The results of this evaluation were that the LLM and domain expert agreed 76% of the time (Table 2). More important than the absolute accuracy of the experiment was an investigation into the specifics of the LLM failures. With the steep development curve of this emerging technology, accuracy is likely to improve with subsequent language models, improved system performance and more nuanced prompt engineering and tuning. If the current failures follow a pattern then this is something that can be addressed logically. If failures, however, are apparently random and hallucinogenic then this may

Total results evaluated	103
Instances where LLM and domain expert agree	78
Agreement rate	75.7%

Table 2: Expert evaluation

be problematic for improving performance and may offer unacceptable outputs in production-ready systems.

After some analysis of the LLM failures, it was noted that most of the failures share a common theme. Some comments from the evaluation:

> ‘This doesn’t sound like a childhood experience to me. The protagonists are merely described as young, and the word ’childish’ is used (suggesting they’re actually young adults).’
>
> ‘Very borderline - the term ’girls’ doesn’t really indicate children here I think.’
>
> ‘No - this is simply a singing voice with a youthful quality.’
>
> ‘Could be a childhood experience but I don’t think there’s much in the text to support this.’

The comments above, and more from the full evaluation, circle mainly around cases where the LLM has been over-optimistic in its analysis, often involving phrases referring to when the subject was young or when they were with their parents. The LLM has interpreted these as instances of childhood but there is often a lack of evidence to say this is definitely the case.

1.4 Discussion

The experiment’s initial results showed the challenges of LLMs in thematic text classification, with notable discrepancies between the LLM and domain expert. The LLM’s propensity to flag both overlooked and spurious entries underscores the difficulties inherent in exhaustive human categorization and the machine’s pattern-recognition capabilities. Subsequent adjustments to the LLM’s prompt, aligned with the CHILD project use case, enhanced classification accuracy, indicating the effectiveness of scenario-based prompting in improving LLM output.

The more focused second attempt revealed a pattern in some of the LLM’s misclassifications: a tendency to overgeneralise childhood indicators. This appears to be a consistent overreach, interpreting any youthful reference or familial mention as a childhood experience. The human expert’s review illuminated this over-optimistic bias, offering clear examples where the LLM conflated youthfulness with childhood.

These findings are promising for future LLM applications; errors followed a discernible theme rather than being completely random, suggesting that with refined training and prompt engineering, the LLM’s discernment could be significantly improved.

1.5 Future work and expansion

The promising results of this small study lay the groundwork for broader applications of LLMs in thematic text classification. A direction for future research is to investigate to what extent this approach can be generalised. Can LLMs be effectively adapted to classify texts across arbitrary themes with minimal human intervention?

Building on the current findings, subsequent studies will aim to refine LLM prompting strategies and discover whether a framework can be developed to speed up the process of deploying new thematic classifiers on demand.

Another approach would be to explore the integration of LLMs with knowledge graphs to enhance thematic classification. By leveraging structured knowledge, LLMs could gain a deeper understanding of themes and their interconnections within and across texts, which may improve classification accuracy.