

Applying Natural Language Processing (NLP) for the Automated Classification of Public Consultation Responses

Background
Public consultation plays a vital role in governance however the volume of free-text responses makes it challenging for analysts to quickly extract insights. This project explores the integration of a GPT-4 model into a typical data ingestion pipeline. The pipeline was developed using FME, GPT-4 on Azure Open AI and Power BI for interactive visualisation of the data outputs.

Aim
To integrate NLP using a LLM into the organisation’s data ingestion pipeline to automate the classification of free text survey responses, enabling immediate visualisation in a Power BI dashboard, with manual review options for the project team.

- Objectives**
1. Research and evaluate NLP techniques suitable for text classification and theme extraction.
 2. Design a pipeline that automatically processes, classifies, and stores survey feedback using NLP.
 3. Reduce analyst workload by automating the classification of consultation responses
 4. View results in Power BI for real time display of classified themes as they come in.
 5. Validate the accuracy and performance of the NLP classifications against historic human classified consultation data.

Methodology
The framework guiding this project is CRISP-DM [1] with six key phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

Technology Choices
Azure OpenAI (GPT-4): selected for secure integration and strong language understanding; practical for a dynamic, multi topic consultation pipeline.
FME: widely used within the organisation, supports triggers/scheduling and API calls for automation.
SQL Server & Power BI: robustness, security, and interactive visualisation.
Prompting: The final theme classification prompt was developed using Microsoft’s GCSE framework - Goal, Context, Source, and Expectation.

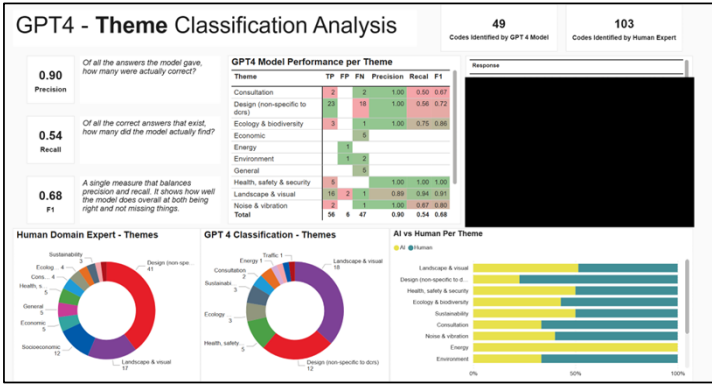
Design and Implementation
Here is a summary of what happens in the data pipeline:

1. Data preparation – responses, themes, and topics are prepared for processing.
2. Theme classification – GPT-4 assigns one or more themes to each response; results are unpacked into tables.
3. Topic classification – filtered by assigned themes, GPT-4 assigns specific topics; results unpacked into tables.
4. Output storage – final results written securely to SQL Server for analysis in Power BI.

- Results**
- **Theme** classification: F1 = 0.68 (good performance, fewer identified compared to human expert).
 - **Topic** classification: F1 = 0.37 (close in numbers to human coding, but lower accuracy).

User Feedback SUS Questionnaire [2]
A questionnaire was distributed to users of the dashboard. Here is a summary of the results:

- 12 responses with positive feedback.
- Easy to use and the functions are well integrated.
- F1 scores valued for transparency of model performance.
- AI seen as assistant, not replacement.



Conclusion
LLMs can deliver benefits for consultation analysis, but AI should be used as an assistant rather than replace expert judgement.

- Adopt hybrid approaches with human reviewers remaining responsible for validation.
- Improve data quality and representation, flagging under-represented groups.
- Increase transparency and interpretability by including confidence scores and example justifications.

References
[1] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000, pp. 29–39.
[2] J. Brooke, "SUS: A 'quick and dirty' usability scale," in Usability Evaluation in Industry, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London: Taylor & Francis, 1996, pp. 189–194.