

CW1 Assignment

Monday, July 28, 2025 9:24 AM

Changemakers: Creativity and Value Creation

Module Code: ACCA5033

Assignment 1: Innovation & Research

Word Count: Approx. 2500

Submission Format: OneNote Export (PDF)

Submission Deadline: August 2025

Project Title:

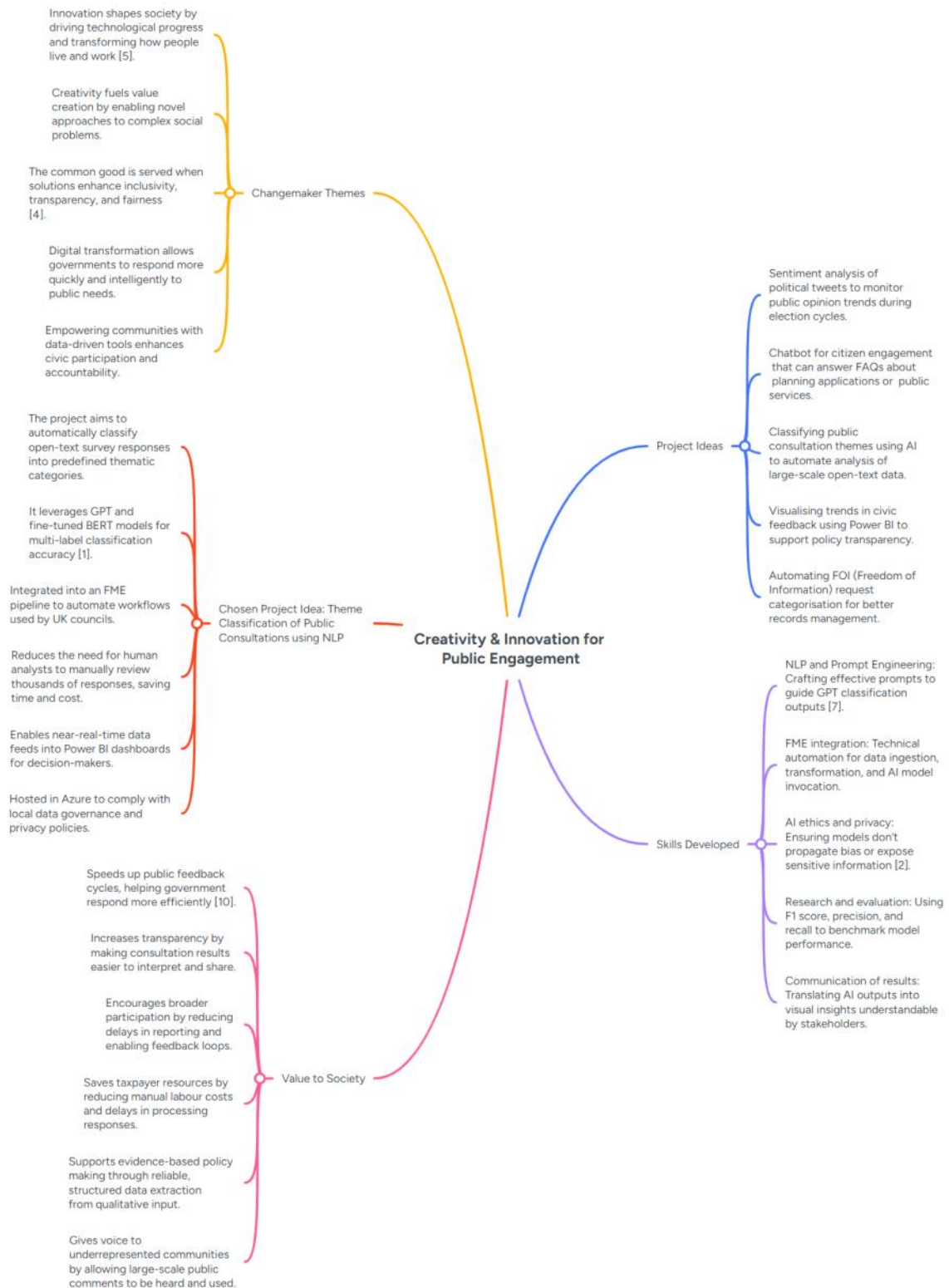
Automating Public Consultation Theme Classification using GPT and FME

Student Name: Rhys Davies

Student ID Number: 2008896

Mind Map – Creativity Innovation for Public Engagement

Monday, July 28, 2025 9:24 AM



Aim & Objectives and Research Area

Monday, July 28, 2025 9:26 AM

Project Aim

I intend to automate the categorisation of open-text feedback in public consultations with a AI two-step oriented workflow within an FME pipeline. Leveraging GPT models hosted securely on Azure, my idea is to alleviate some of the manual workload required when performing thematic analysis, while also enabling the results to be funnelled directly into Power BI dashboards for real-time insight during public sector decision-making.

SMART Objectives

1. To create the two-stage GPT-driven NLP pipeline for grasping both overarching themes and narrower subtopics in each response [1].
2. To include this classification process in a running FME Workbench flow that is performing handling consultation data by councils.
3. To format the outputs in a form that is compatible with Power BI dashboards through which automatic visual reporting to consultation teams could be implemented.
4. To test the accuracy and the performance of the model using an annotated dataset based on precision, recall and F1 score [3].
5. That the approach is ethically, legally and practically sound with regards to reducing hallucinations, minimising bias [2] and maintaining all data processing within the Azure environment.

Research Philosophy

I've taken a pragmatic research philosophy [6] because my project is about addressing real-world issues via repeated actions of exploring, designing, and assessing. But I'm not just trying to figure out in theory how these big language models work, I'm working on how to make them valuable in a practical applied setting: public consultation analysis.

My approach comprises of quant method (such as F1 score) and qualitative method (such as clarity of the output, usefulness to the stakeholder) Benight pragmatism, that enables me to dip into both approaches and move back and forth between them according to what is most effective at any point in time. It's also great for mitigating technologies like GPT, where success depends as much on prompt design and context [7] as it does on raw model architecture.

Research Tools & Databases

To help me develop and validate this project I used a few of these tools and sources: IEEE Xplore and Google Scholar assisted with the papers for technical papers on multi-label classification, fine-tuned BERT [3] and large language models. I searched Scopus and the ACM Digital Library for ethical considerations in the deployment of AI and practical uses of civic tech. OpenAI's documentation was invaluable for learning about what GPT could do, when it shouldn't be used and best practices [1].

I also reviewed gov.uk consultation datasets and local authority guidance to develop the real-use case for my system.

Relevance to My Programme

This project draws on several threads from my computing coursework: cloud development, AI integration, automated workflows and data visualisation. It's also training me for the types of work I do in my company, applying machine learning in actual, publicly facing products, particularly in scenarios where transparency, ethics [4] and performance matter. The project has been influenced not simply by academic exploration, but by my own experimentation, review and gradual improvements across time. I've begun to iterate on it based on how the models are behaving, how it's being visualised on dashboards, and what real-world user feedback tells me.

Research Strategies

Monday, July 28, 2025 9:26 AM

Q1. Validity, Reliability, Generalisability and Ethical Considerations

Validity is ensured by training the GPT model to focus on real-world sensibility: responses are matched only with known themes and subthemes distilled from historical consultation analysis. This restriction is designed to keep classifications meaningful and based on stakeholder intention (i.e. not artificial or fictitious).

The reliability is strengthened with the help of consistency checks, that is, each step in the classification pipeline is executed multiple times for the same inputs to gauge output consistency. Moreover, testing with same GPT version (eg, GPT-4 on Azure) across testing stages reduces the variance from model versions.

Generalisability is dealt with by framing the prompts and theme taxonomy to be applicable to different consultation types (e.g. transport, housing, budget). Using historic datasets from multiple U.K. councils makes the project a rich and diverse training and evaluation ground. The approach itself, although the "tuning" can be problem-specific, is easily transportable to other FME-based councils.

Ethical considerations are central to the project:

- GPT is hosted via **Azure OpenAI** to comply with internal data governance requirements and avoid sending public data to external servers.
- Prompts are designed to minimise **hallucinations** [1](false or speculative outputs) by restricting the context given to the model.
- The second GPT step is only triggered **after theme detection**, which acts as a gatekeeper and reduces over-classification.
- Bias is monitored through manual review of output patterns, particularly for sensitive topics like race, accessibility, and inequality [2].

Additionally, because the data involves public comments, personally identifiable information (PII) is anonymised or excluded where necessary.

Q2. Research Strategy and Pilot Method

This project uses a **mixed methods approach** grounded in pragmatism [6]:

- **Quantitative methods:** Automated model outputs are compared to manually labelled datasets using classification metrics:
- **F1 Score** (harmonic mean of precision and recall)
- **Precision** (relevance of assigned labels)
- **Recall** (coverage of actual themes)
- **Qualitative methods:** Prompt outputs are reviewed manually by domain experts to assess clarity, topic appropriateness, and usefulness for dashboard reporting. A scoring sheet is used to track subjective quality indicators such as response relevance and usefulness in a policy context.

Pilot Process

A pilot test has already been conducted with a sample dataset of 100 open-text responses from a recent active travel consultation. Key elements of the pilot included:

- Testing both GPT steps independently to measure accuracy loss between stage one (theme detection) and stage two (subtopic classification).
- Collecting stakeholder feedback on the quality of outputs, especially where responses were vague or multi-thematic.
- Iteratively adjusting the prompt format — including examples and context instructions — to improve performance.

Q3. Quantitative and Qualitative Analysis Methods

Quantitative Methods:

- **F1 score:** The primary evaluation metric, which combines the precision and recall.
- **Support-weighted scores:** Used to normalise scores for the class imbalanced (some themes occur much more frequently than others).
- **Performance comparisons** between GPT and fine-tuned BERT [3] will also be included if time permits [8].

Qualitative Methods:

- **Manual review of classifications** against domain expert interpretations [12].
- **Prompt analysis:** Identifying which prompt elements led to hallucination or underperformance.
- **Stakeholder feedback sessions:** Focused on dashboard usability and output interpretability.

Together, these analysis techniques ensure the system is not only statistically robust, but also practically valuable and ethically sound for long-term deployment in government workflows.

References

Tuesday, July 29, 2025 12:04 PM

- [1] OpenAI, "GPT-4 Technical Report," 2023. [Online]. Available: <https://openai.com>
- [2] M. Morstatter et al., "Social Bias in Large Language Models," *AAAI Conference on Artificial Intelligence*, 2023.
- [3] J. Zhou et al., "Multi-label Text Classification with BERT," *Neural Computation*, vol. 32, no. 5, pp. 1120–1145, 2020.
- [4] E. Binns et al., "Fairness and Accountability in Machine Learning," *Communications of the ACM*, vol. 61, no. 6, pp. 64–73, 2018.
- [5] E. M. Rogers, *Diffusion of Innovations*, 5th ed., New York: Free Press, 2003.
- [6] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 8th ed., Pearson Education, 2019.
- [7] D. H. Park et al., "Prompt Engineering Techniques for Improving GPT Output," *arXiv preprint arXiv:2302.11382*, 2023.
- [8] A. Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [9] K. Crawford, "Artificial Intelligence's White Guy Problem," *The New York Times*, June 25, 2016.
- [10] Gov.uk, "Consultation Principles: Guidance," UK Government, 2018. [Online]. Available: <https://www.gov.uk/government/publications/consultation-principles-guidance>
- [11] A. McKinney, "FME and Power BI Integration for Civic Data Reporting," *Safe Software Blog*, 2023.