NMEMS: THE AUTOMATED MAPPING PROJECT

AUTHORS
AISHA OGUNSOLA

SPECIAL THANKS TO:

- Lisa Hopcroft - Senior Technical Consultant, NHS NSS
- Simon Rogers - Senior Technical Consultant, NHS NSS

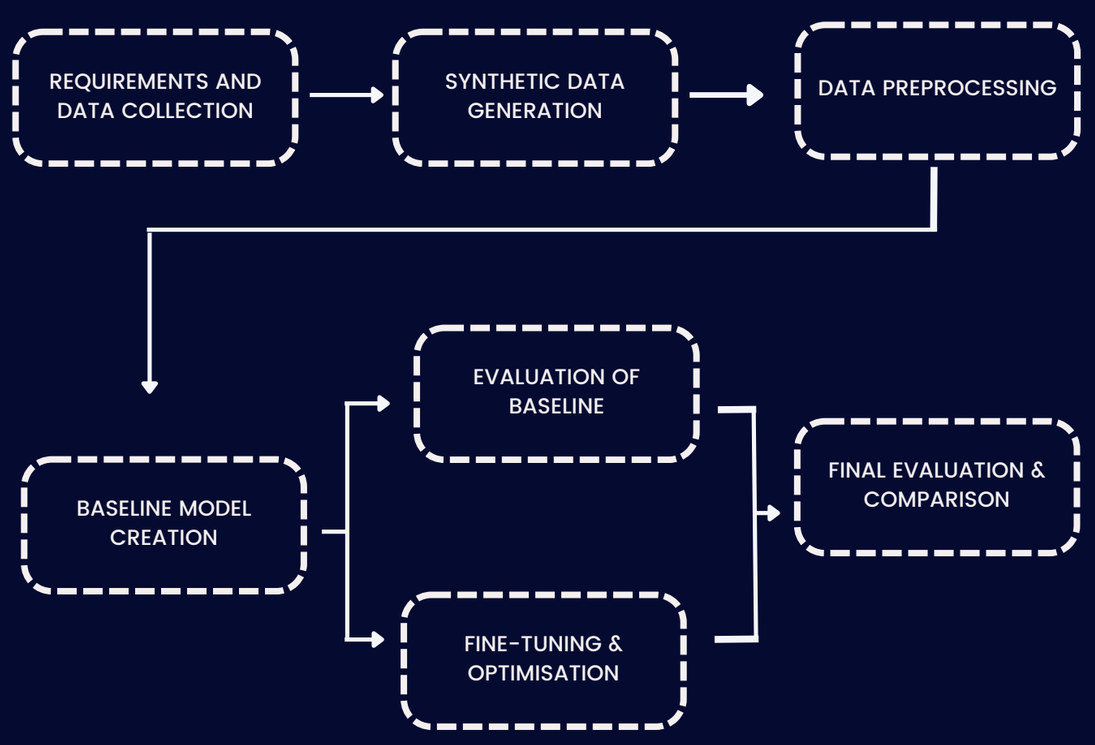
BACKGROUND

- Each health board in Scotland maintains its own equipment database → inconsistent naming.
- During COVID, the urgent need to locate ventilators showed the importance of a national harmonised database.
- AI, particularly large language models (LLMs), can support the unification of local names under a single national standard.

OBJECTIVE

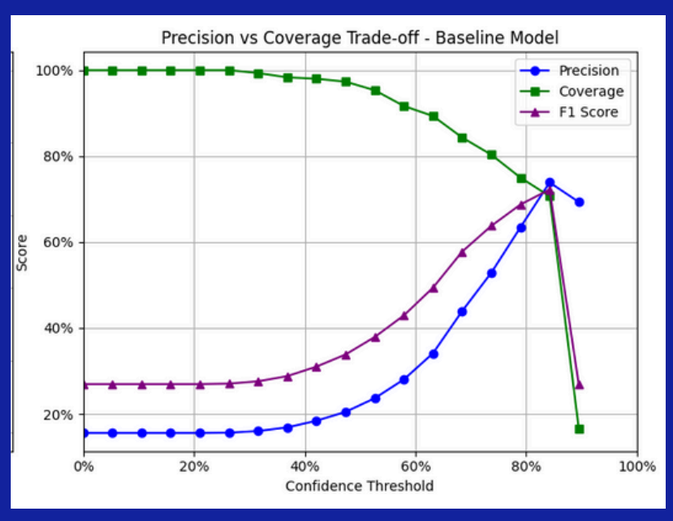
- Build AI tools to map local names → national names.
- Provide a confidence score for each match.
- Benchmark against a manually curated gold-standard dataset.
- Fine-tune and optimise models for accuracy and reliability.
- Compare different models to identify the best-performing approach.

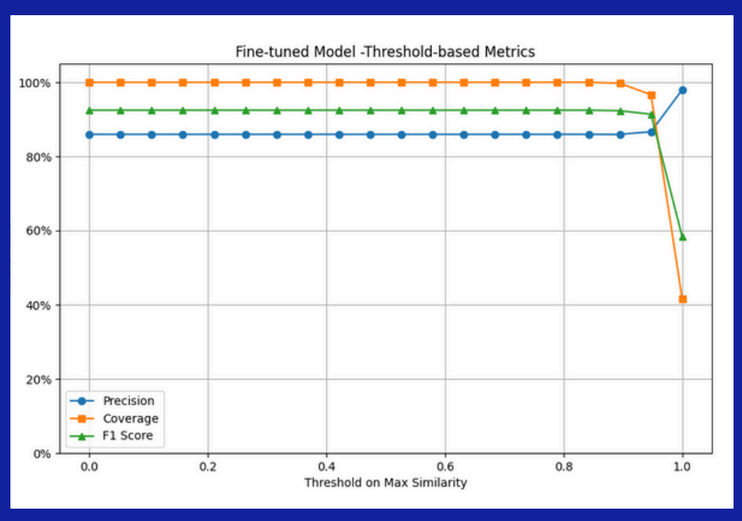
METHODOLOGY



ANALYSIS

A Sentence Transformer based on BERT measured similarity between local and national equipment names, with standard BERT serving as the baseline. Dataset augmentation generated synthetic pairs by swapping and reordering words to expand coverage. Pre-processing, including punctuation removal, ensured clean embeddings. Fine-tuning adapted the model to the specific medical equipment context, improving accuracy, precision, and top-k recall. Evaluation metrics, including accuracy, precision, F1, and top-k recall, assessed prediction reliability and ranking quality.





RESULTS

The fine-tuned model outperforms the baseline across most evaluation dimensions:

- Higher and more stable precision
- Maintained coverage until very strict thresholds
- Improved top-k retrieval performance

CONCLUSION

The fine-tuned model demonstrates strong potential to standardise category mapping, reduce manual effort, and improve data quality across NHS health boards. Careful deployment with feedback loops and ongoing monitoring will maximise reliability and usability in operational settings.

GLOSSARY OF KEY TERMS

- Synthetic Data Generation: Creating artificial data using algorithms to simulate real-world data. This helps train models when actual data is limited or sensitive.
- Baseline Model: A simple starting version of a machine learning model used for comparison with improved versions.
- Fine-Tuning & Optimisation: Adjusting and improving a model to make it more accurate and efficient for a specific task.
- Data Pre-processing: Cleaning and preparing raw data (e.g., removing punctuation or formatting) before using it in a model.
- Sentence Transformer / BERT: Advanced AI models that understand and compare the meaning of sentences. BERT is a popular model used for language tasks.
- Embeddings: Numerical representations of words or sentences that help AI understand their meaning.

GLOSSARY OF KEY TERMS

- Top-k Recall: A way to measure how often the correct answer appears in the top few results returned by the model.
- Evaluation Metrics (Accuracy, Precision, F1 Score): Different ways to measure how well a model performs. For example:
  - Accuracy: How often it's right.
  - Precision: How often its positive predictions are correct.
  - F1 Score: A balance between precision and recall.
- Gold-standard Dataset: A high-quality, manually checked set of data used as a benchmark to test models.
- Confidence Score: A number showing how sure the model is about its prediction.
- Category Mapping: Linking different names or labels (e.g., local vs. national equipment names) to a single standard category.
- Feedback Loops: Using user input or results to continuously improve the system.