

Leveraging Artificial Intelligence to Drive Timely Improvements in Patient Experience Feedback: Algorithm Validation

Mustafa Khanbhai, Catalina Carenzo, Sarindi Aryasinghe, Dave Manton, Erik Mayer

Submitted to: JMIR Medical Informatics
on: May 24, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	40
Figures	41
Figure 1.....	42
Figure 2.....	43
Figure 3.....	44
Figure 4.....	45
Figure 5.....	46
Multimedia Appendixes	47
Multimedia Appendix 0.....	48

Leveraging Artificial Intelligence to Drive Timely Improvements in Patient Experience Feedback: Algorithm Validation

Mustafa Khanbhai¹ MD, PhD; Catalina Carenzo¹ MSc; Sarindi Aryasinghe¹ BSc; Dave Manton² BSc; Erik Mayer¹ MD, PhD

¹ Imperial Clinical Analytics, Research and Evaluation (iCARE), Imperial College London London GB

² NIHR Imperial Biomedical Research Centre Imperial College London London GB

Corresponding Author:

Mustafa Khanbhai MD, PhD

Imperial Clinical Analytics, Research and Evaluation (iCARE),
Imperial College London
The Bays, Entrance 2
South Wharf Road, St Mary's Hospital
London
GB

Abstract

Background: Understanding and improving patient care is pivotal for healthcare providers. With increasing volumes of the Friends and Family Test (FFT) data in England, manual analysis of this patient feedback poses challenges for many healthcare organisations. This underscores the importance of automated text analysis, particularly in predicting sentiments and themes in real-time.

Objective: Leveraging machine learning and natural language processing, this study explores the utility of a supervised algorithm to systematically test and refine the algorithm's cross-contextual performance in diverse healthcare settings, addressing variations in population characteristics, geographical locations, and care settings, ultimately driving improvements based on patient feedback.

Methods: The text analytics algorithm initially developed in a large acute Trust in London was further tested in nine healthcare organisations with diverse care settings across England. These Trusts varied in technical capacity and resource, population demographics, and FFT free text datasets. Testing and validation of the algorithm was performed including manual coding of subset of retrospective comments. Technical infrastructure was optimised including coding environments and packages for algorithm testing and deployment. The algorithm was iteratively trained using bag of words from anonymised data, tailored to accommodate contextual variations, and tested for change in algorithm performance whilst simultaneously rectifying issues identified.

Results: The algorithm demonstrated satisfactory overall accuracy (>75%) in predicting themes and sentiments embedded within free-text responses across a variety of care settings and population demographics. While the algorithm yielded strong and reusable models in relatively stable environments, such as adult inpatient care settings, the initial accuracy was notably lower in organizations providing services such as paediatrics and mental health. However, the accuracy of our algorithm significantly improved when individual Trust coding templates were applied. Thematic saturation was reached after the fifth organisation was recruited, and no further coding was required for the last four organisations. Subsequently, a framework and pipeline for deployment of the algorithm were developed to provide standardised approach for implementation and analysis of FFT free text, ensuring ease of use.

Conclusions: This study represents a significant step forward in leveraging free-text FFT data for valuable insights in diverse healthcare settings through the testing and development of a robust supervised learning text analytics algorithm. The disparity in some care settings was anticipated, given that the lexicon and phraseology used was inherently different from those prevalent in adult inpatient care (where the algorithm was developed). However, these challenges were addressed with further coding and testing. This approach enhanced the accuracy and reliability of the algorithm, encouraged inter- and intra-organisational collaboration, and shared learning.

(JMIR Preprints 24/05/2024:60900)

DOI: <https://doi.org/10.2196/preprints.60900>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Leveraging Artificial Intelligence to Drive Timely Improvements in Patient Experience Feedback: Algorithm Validation

Mustafa Khanbhai^{1 2}, Catalina Carenzo^{1 2}, Sarindi Aryasinghe^{1 2}, Dave Manton², Erik Mayer^{1 2}

¹ Imperial Clinical Analytics, Research and Evaluation (iCARE), Department of Surgery & Cancer, Imperial College London, UK

² NIHR Imperial Biomedical Research Centre, UK

Abstract

Background

Understanding and improving patient care is pivotal for healthcare providers. With increasing volumes of the Friends and Family Test (FFT) data in England, manual analysis of this patient feedback poses challenges for many healthcare organisations. This underscores the importance of automated text analysis, particularly in predicting sentiments and themes in real-time.

Objective

Leveraging machine learning and natural language processing, this study explores the utility of a supervised algorithm to systematically test and refine the algorithm's cross-contextual performance in diverse healthcare settings, addressing variations in population characteristics, geographical locations, and care settings, ultimately driving improvements based on patient feedback.

Methods

The text analytics algorithm initially developed in a large acute Trust in London was further tested in nine healthcare organisations with diverse care settings across England. These Trusts varied in technical capacity and resource, population demographics, and FFT free text datasets. Testing and validation of the algorithm was performed including manual coding of subset of retrospective comments. Technical infrastructure was optimised including coding environments and packages for algorithm testing and deployment. The algorithm was iteratively trained using bag of words from anonymised data, tailored to accommodate contextual variations, and tested for change in algorithm performance whilst simultaneously rectifying issues identified.

Results

The algorithm demonstrated satisfactory overall accuracy (>75%) in predicting themes and sentiments embedded within free-text responses across a variety of care settings and population demographics. While the algorithm yielded strong and reusable models in relatively stable environments, such as adult inpatient care settings, the initial accuracy was notably lower in organizations providing services such as paediatrics and mental health.

However, the accuracy of our algorithm significantly improved when individual Trust coding templates were applied. Thematic saturation was reached after the fifth organisation was recruited, and no further coding was required for the last four organisations. Subsequently, a framework and pipeline for deployment of the algorithm were developed to provide standardised approach for implementation and analysis of FFT free text, ensuring ease of use.

Conclusion

This study represents a significant step forward in leveraging free-text FFT data for valuable insights in diverse healthcare settings through the testing and development of a robust supervised learning text analytics algorithm. The disparity in some care settings was anticipated, given that the lexicon and phraseology used was inherently different from those prevalent in adult inpatient care (where the algorithm was developed). However, these challenges were addressed with further coding and testing. This approach enhanced the accuracy and reliability of the algorithm, encouraged inter- and intra-organisational collaboration, and shared learning.

Introduction

Insights into patients experience of their care provision is critical for healthcare providers delivering patient-centred care and recognised as integral to ensuring safe and high-quality care.[1] By the end of 2019, the national survey for patient experiences in the United Kingdom (UK), known as the Friends and Family Test (FFT), had accumulated 75 million pieces of feedback.[2] This number continues to grow by about 1.3 million every month, making it the largest source of patient opinions globally. However, manually analysing this vast amount of data requires a considerable resource, something that many healthcare organisations struggle to provide.[3, 4] In recent years there has been an ever-clearer desire for improvement to arise from patient feedback rather than just focusing on response rates, but most healthcare organisations are constrained from doing so due to lack of analytical expertise.[5]

Automated text-analysis by Artificial Intelligence (AI) and its component Natural Language Processing (NLP) predicts sentiments and themes within free-text responses in real-time and provides a useful opportunity for better integrating patient experience feedback into everyday healthcare delivery.[4] NLP algorithms, offer a practical solution to decode the 'why' behind patient narratives and through their ability to process large volumes of data accurately these patient-derived insights can inform 'business as usual' within healthcare organisations.[6, 7] Recent advances in natural language processing (NLP), particularly the development of Large Language Models (LLMs) and transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) [8] and GPT (Generative Pre-trained Transformers) [9], have significantly improved the analysis of patient feedback in healthcare. These models replace traditional hand-coded, framework-based categorization [10] methods by automating the classification of concerns and providing decision-makers with real-time, comprehensive summaries of emerging themes. LLMs, through Multi-label Text Classification (MLTC) [11], enable the extraction of deeper insights from patient experience comments, enhancing the understanding of the patient journey and informing actions that can improve care. Transformer-based models like BERT [8, 12] and GPT excel at capturing complex linguistic patterns, context, ambiguity, and sentiment, making them particularly well-suited for the nuanced task of sentiment analysis in healthcare. These advancements allow for more accurate, context-aware analysis of patient

feedback, driving more actionable insights and supporting continuous improvements in patient care.

BERT and GPT models are large and resource intensive.[13] They require significant computational power and memory, which may not be feasible in all healthcare settings, especially those with limited IT infrastructure. This can make them less practical for real-time applications or for use in smaller organizations with fewer resources. Due to their size and complexity, transformer models can be slower to process compared to more lightweight models like support vector machines (SVMs). In real-time healthcare applications where speed is crucial for timely interventions, this delay may be a significant drawback.

We have previously demonstrated the benefit of applying NLP using SVM to free-text FFT data to deliver patient driven quality improvement in a single institution.[14] This study demonstrates the broader implementation and iteration of this algorithm in different hospital setting in England that use FFT to capture patient feedback,

Methods

The implementation of the FFT free-text algorithm demands thorough testing to ensure robustness, accuracy, and applicability. To iterate the algorithm for deployment in other healthcare services, we first selected healthcare organisations that would incorporate geographical variables, patient demographic variables, and different care settings, i.e., adult inpatient, primary care, community, paediatrics, and mental health.

The objectives of this study were to identify technical capabilities for algorithm deployment in the included healthcare settings, identify variation in user needs across different service settings and geographies, and iterate the algorithmic process and deploy a reproducible algorithm that can be implemented in different services.

Trust recruitment

A steering group consisting of lay representatives and patient experience experts was formed to guide the strategic direction of this study and help in recruitment through various patient experience networks including the National Insight Network set up by NHS England, the Q Community established by the Health Foundation, Advancing Quality Alliance, and the Heads of Patient Experience Network. A FFT review capability questionnaire (Appendix 1) was sent to interested organisations. This allowed the study team to identify organisations based on three key metrics, patient experience engagement, digital readiness/information technology infrastructure, and quality improvement involvement. To ensure that the selected Trusts possess the necessary infrastructural and organisational qualities to contribute effectively to the research objectives while adhering to data regulatory standards, a capacity and capability assessment was required. Only those Trusts that were able to confirm capacity and capability through their Research and Development department were finally recruited, as shown in Figure 1.



© Imperial College Healthcare NHS Trust 2024. All rights reserved. For licensing enquiries contact nhsinfo@imperial.ac.uk

Figure 1. An overview of the geographical distribution of the nine recruited Trusts.

Ethical Considerations

This study received Health Research Authority and Health and Care Research Wales approval, 20/HRA/5924. No patient data was collected, and Trusts were anonymised in the analysis. Written consent was sought from members of the steering group and individual hospital Trust project team, and with further written consent their contribution has been acknowledged in the acknowledgements section.

Pre-implementation

IT scoping tool

To work collaboratively, we identified the needs of all participating organisations from a

patient experience digital readiness standpoint. This was facilitated through previous work by the study team funded by NHS England Insight and Feedback Team [15] where a scoping tool was developed to identify key contacts, FFT data collection, structured query language (SQL) Server database, handling and management, infrastructure capacity i.e., on-premise data centre or cloud computing, integrated development environment and coding language expertise, and data visualisation software (Appendix 2).

Data access

To ensure that participating organisations were fully supported by the project team a Data Protection Impact Assessment (DPIA) was completed with each organisation. DPIA is a systematic process designed to identify and minimize the associated data protection risks. The purpose of a DPIA in our study was to assess the impact of processing data and to ensure that the principles of data protection, such as those outlined in regulations like the General Data Protection Regulation were adhered to.

Adaptation of coding script

To achieve transferability of the support vector machine (SVM) supervised machine learning (ML) algorithm and test against different service settings and NHS provider organisations, the functionalities previously developed [7] were re-coded using Python (Python Software Foundation) given the popularity of using this coding platform in most NHS Trusts. Python is an interpreted high-level general-purpose programming language. The model was developed with Python 3.8.10 with the following free Python libraries; scikit-learn $\geq$ 0.22.1 [16] (a machine learning library that provides relevant machine learning packages required for classification), Natural Language Toolkit (NLTK) $\geq$ 3.4.4 (25) (a suite that contains libraries and programs to make machines understand human language), and Pandas $\geq$ 0.25.3. (26) (a library for data manipulation and analysis). These requirements were then translated into a requirements.txt file for organisations to deploy the algorithm. SVM was used as it is the most common classifier widely used for document classification, which consistently yield good classification performance.[10]

Updating training coding template

All FFT free-text comments were coded based on the NHS Patient Experience Framework (Appendix 3). This framework incorporates eight key themes that outline those elements which are critical to the patients' experience of NHS services.[17] In the previous work [7], three additional themes were added; 'Unclassified' (too many typographical errors to

discern meaning), 'General' (e.g., NHS is great, everything), and 'Staff' (comments relating to staff). From the NHS England funded work [15] it was observed there was an overlap of comments between theme 'Staff' and other themes such as 'Respect for patient-centred values, preferences, and expressed needs', 'Emotional support', and 'Physical comfort'. This is because often a staff member was involved in the delivery of the care as it related to the other themes. The theme 'Staff' was removed, and all comments coded as 'Staff' were recoded to develop a concise coding template. Therefore, 10 themes were selected for final analysis and a coding pack (Appendix 4) was developed with detailed guidance. Furthermore, we iterated the initial algorithm from our previous work which was developed using two follow-up free-text questions, i.e., 'What did we do well?' and 'What could we do better?', by combining and shuffling the training data from both questions to develop a streamlined SVM supervised machine learning algorithm for theme and sentiment prediction.

Implementation

A pipeline was constructed for the SVM model in Python for text classification and sentiment analysis, consisting of several steps, including Text Pre-processing, Feature Engineering, term frequency-inverse document frequency (TFIDF) transformation, and model training and testing.

FFT data retrieval

SQL Server was used to both extract the FFT data and supplementary free text for algorithm imputation, and to store the output of the algorithm. Pyodbc >= 5.0.1 Python library was required to connect to the individual organisations SQL server. SQL's combination of ease of use, flexibility and capabilities make it useful for storing, managing, querying, and analysing datasets and is widely used in healthcare organisations.

Establishing ground truth

In supervised learning algorithms, ground truth data is critical to training and updating an algorithm.[10] Therefore, more annotated data was required to improve the performance of the algorithm. The coding pack (Appendix 4) developed was provided to the participating

organisations patient experience teams to make coding more feasible. The coders were a combination of patient experience team, clinical staff and Trust lay representatives. Two coders from the team completed manual coding individually using 500 stratified sample to calculate inter-annotator reliability and understand the interpretation of themes and sentiment. Disagreements in coding were reviewed, and theme and sentiment were finalised to create a master coding template in .csv file. This template was populated with data from all nine organisations.

Pre-processing

Pre-processing of textual data is the first essential step in the processing of text and has been proven to improve text classification models' performance by standardising the text before it is presented to the classification algorithm.[10] NLTK for Python was used for pre-processing and data analysis which includes tokenization of the comments and removal of stop words (e.g., the, was, is, etc.), punctuation, numbers, and special characters from the comments. The resulting words were represented as a bag of words (BoW) or corpus. Given that the average length of FFT free-text comments tends to be short, and the context is domain specific, i.e. patient experience, this bag of words approach was used rather than Word Embedding.

Feature engineering

Feature engineering is the process of transforming data into features/attributes that better represent the underlying structure of the data.[10] To identify the key features of the data, we used a statistical measure TFIDF (Term Frequency Inverse Document Frequency), used in the fields of information retrieval (IR) [18] and machine learning, that can quantify the importance or relevance of words in a comment amongst a collection of comments.[19]

Determining accuracy

To understand if further coding was required, the accuracy for theme and sentiment was calculated using 500 stratified sample of manually coded comments from each of the individual organisations. The accuracy is defined as a metric used in classification to measure the percentage of accurate predictions.[10] It is also defined as the ratio of number of correct predictions by the algorithm to the total number of predictions. In other

words, ratio of true positives (TP) and true negatives (TN) to all positive and negative observations (where FP and FN are false positives and false negatives respectively): $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$. The acceptable threshold for overall accuracy for theme and sentiment for a supervised algorithm using patient feedback was determined at 75%. The 75% accuracy threshold for theme and sentiment classification was established based on the work of the study group [10], and previous studies in the literature.[20, 21] Specifically, the threshold reflects the balance between achieving an acceptable level of accuracy and the complexity of interpreting nuanced patient feedback using NLP and ML techniques. Therefore, the 75% accuracy benchmark was identified as a reasonable trade-off between high-quality model performance and the inherent challenges in processing and classifying subjective feedback from patients.

The accuracies yielded by the models on different ranges of hyperparameters were compared. For SVM, both L1 and L2 regularization, variance tolerance values for the stopping criteria, and different values for the inverse of regularization strength, C, were compared. As regularization has a known role in reducing overfitting, the hyperparameters were chosen for tuning to identify the amount and type of regularization and early stopping criteria that are suitable for the classification tasks. After this, the optimal hyperparameters found by GridSearchCV from the scikit-learn library. A 10-fold cross-validation was used to conduct a search to select the best hyperparameter combinations from specified ranges of hyperparameters, which were used to re-train the model on the training set and to predict on the test set.[10] This allowed us to establish the accuracy but also use the coded comments as a template to improve the accuracy by a further round of re-coding if needed until the accuracy was at or above 75%, i.e., reached thematic saturation.

Algorithm deployment

The master coded template data was then serialised using the Python pickle library (28). This module is used for serializing and de-serializing objects into Python format text strings, which are called pickles and can be created by Python code or sent from other Python programs. This Python pickle is easily carried over the network and can be validated without knowing the details of the received object. This enables the preservation and sharing of the ML models, allowing users to reload pre-trained models, significantly reducing the need for lengthy re-training. The files that were stored in the environment were

as follows: Feature_sentiment.pkl, Feature_theme.pkl, Sentiment_classifier.pkl, Theme_classifier.pkl, Tfidftransformer_sentiment.pkl , Tfidftransformer_theme.pkl.

Establishment of a Community of Practice

In the pursuit of ensuring the correct sustainability of the algorithm, a strategic initiative was undertaken through the establishment of a Technical Community of Practice (CoP). This community served as a conduit for regular interaction and collaboration among the leaders responsible for overseeing the algorithmic maintenance across different Trusts. Recognising the complexity and potential challenges associated with deploying advanced algorithms, quarterly meetings were instituted as a core component of the CoP framework. These meetings provided a structured way for the leads of each site to convene, fostering an environment conducive to the exchange of critical insights, best practices, and lessons learned.

Results

Nine Trusts were recruited from a range of care settings, sociodemographic and geographical areas. Trusts were added to the study sequentially, one by one, as part of the recruitment process. None of these organisations had a pre-existing automated process for analysing their FFT free-text responses.

Identifying free-text comments with FFT survey for analysis

In the pre-implementation scoping exercise, it was observed that among the participating organizations, Trust A and D incorporated three follow-up free-text questions in their FFT survey. In contrast, all other Trusts included only one follow-up question (Table 1). The question with the highest number of comments in Trust A was analysed. The organisations provided a variety of services, including paediatrics, inpatient and outpatient, and community and mental health thereby facilitating testing and iteration of the algorithm for use in all care settings. Furthermore, most of the organisations 88.9% (8/9) opted to house and deploy the algorithm on premise and the rest on Cloud 11.1% (1/9).

Table 1. An overview of the variation in FFT free-text data, service setting and environment favoured to deploy the machine learning algorithm in nine NHS organisations in England. Version deployed refers to the version of the iterative addition of words to the BoW model, used to improve model performance.

<i>Trust</i>	<i>Service coverage</i>	<i>Model deployment environment</i>	<i>Number of free-text columns</i>	<i>Free-text Question(s)</i>	<i>Version deployed</i>
A	Community and Mental Health	On Prem	3	<i>Please can you tell us why you gave your answer?^a</i>	Version 0
				<i>Please tell us at least one thing that went well.</i>	
				<i>Please tell us at least one thing we could do better.</i>	
B	Acute and Inpatient	Cloud	1	<i>Please can you tell us the main reason for the score that you have given?</i>	Version 1
C	Acute and Inpatient	On Prem	1	<i>Please can you tell us why you gave your answer and what we could have done better?</i>	Version 2
D	Paediatrics	On Prem	3	<i>FFT rating description^a</i>	Version 3
				<i>Write what you think was good.</i>	
				<i>Write what you think was bad.</i>	
E	GP/Community	On Prem	1	<i>Please can you tell us why you gave your answer?</i>	Version 3
F	Acute and Inpatient	On Prem	1	<i>Can you tell us why you gave that response?</i>	Version 3
G	Acute and Inpatient	On Prem	1	<i>Please can you tell us why you gave your answer and what we could have done better?</i>	Version 3
H	Acute and Inpatient	On Prem	1	<i>What was good about your care, and what could be improved?</i>	Version 3
I	Acute and Inpatient	On Prem	1	<i>Please can you tell us why you gave your answer.</i>	Version 3

^a The free text from this question was used in the algorithm testing given the highest responses compare to the other two questions

Variation in patient experience themes

To understand the variation in themes, if any, of FFT free-text responses in the nine organisations, we looked at the count of comments from the FFT free-text responses based on the NHS patient experience framework. This was extracted from the output after initial model deployment for the period of January 2022 to January 2023 as demonstrated in Table 2. The highest count of themes was 'respect for patient centred values' followed by 'general'. The themes with the lowest count were 'welcoming the involvement of friends and family' followed by 'transition and continuity'. There were no demonstratable trends in the other patient experience framework themes.

Table 2. Distribution of count of themes based on the NHS Patient Experience Framework from nine NHS organisations in England, extracted from the output of the algorithm.

<i>Trust</i>	<i>% of FFT comments provided per NHS Patient Experience Framework themes</i>									
	0	1	2	3	4	5	6	7	8	9
<i>A</i>	9	26	4	10	6	6	0	2	24	12
<i>B</i>	8	35	2	14	6	5	2	2	11	16
<i>C</i>	6	23	3	13	8	10	1	2	8	26
<i>D</i>	15	17	9	10	6	5	0	2	16	20
<i>E</i>	22	25	2	11	5	5	0	1	18	12
<i>F</i>	2	43	1	12	4	7	1	1	8	21
<i>G</i>	6	23	3	13	8	10	1	2	8	25
<i>H</i>	7	31	3	7	13	5	2	1	7	25
<i>I</i>	1	25	6	8	9	11	2	3	11	22

0 Unclassified, 1 Respect for patient-centred values, 2 Coordination & integration of care, 3 Information and communication, 4 Physical comfort, 5 Emotional support, 6 Involvement of family and friends, 7 Transition and continuity, 8 Access to care, 9 General.

Establishing ground truth

Inter-rater agreement was determined on the manual coding, and the coding template was updated to check accuracy and improve the model by acquiring an updated corpus from the BoW and feature extraction. The inter-rater agreement was better for sentiment compared to theme. There was almost perfect agreement (0.81–1.00) seen in 7 organisations. Trust A was the only organisations with a substantial (0.61–0.80) agreement with a Cohen's kappa of 0.79 (Table 3).

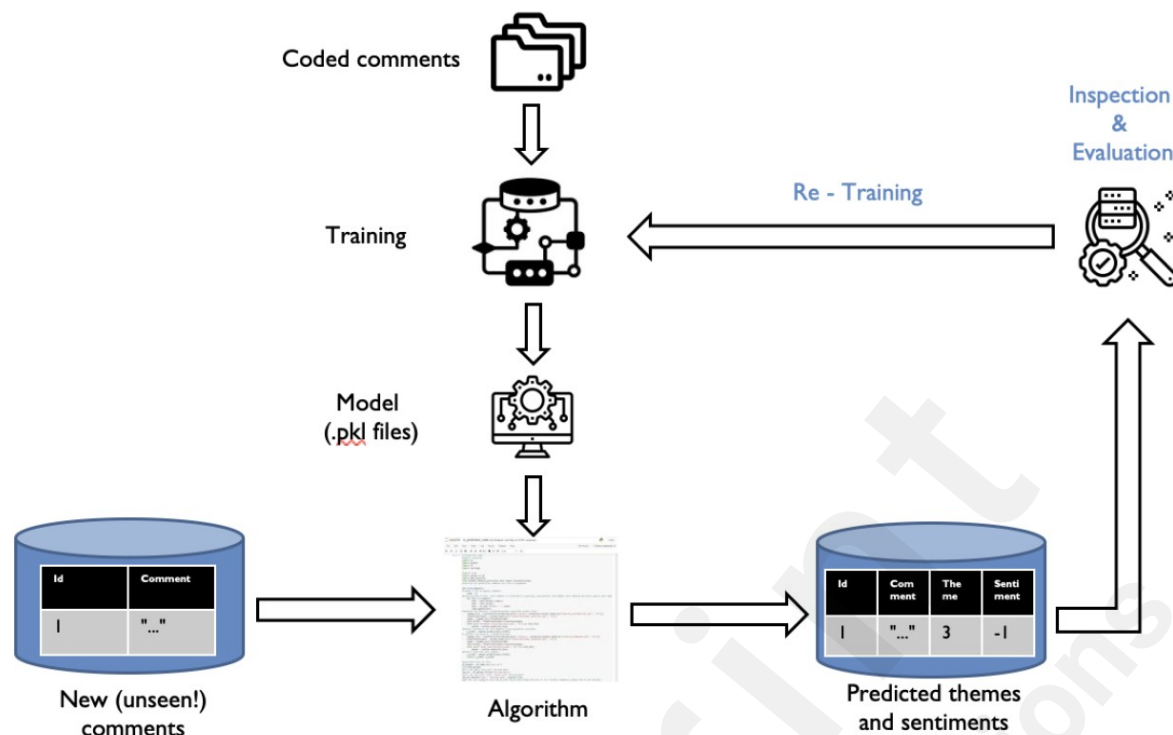
Table 3. Assessing the consistency of theme and sentiment ratings, and final accuracy via

updated coding template.

<i>Trust</i>	<i>Category</i>		<i>Interrater reliability / Cohen's Kappa</i>	<i>Accuracy before re-coding (%)</i>	<i>Final accuracy (%)</i>
A	Community and Mental Health	Theme	0.79	57	76
		Sentiment	0.82	76	82
B	Acute and Inpatient	Theme	0.84	69	78
		Sentiment	1	87	91
C	Acute and Inpatient	Theme	0.85	84	90
		Sentiment	0.91	87	94
D	Paediatrics	Theme	0.89	67	82
		Sentiment	1	72	82
F	GP/Community	Theme	0.80	64	77
		Sentiment	0.82	69	80
E	Acute and Inpatient	Theme	0.96	N/A	>75
		Sentiment	1	N/A	>75
G	Acute and Inpatient	Theme	0.83	N/A	>75
		Sentiment	0.85	N/A	>75
H	Acute and Inpatient	Theme	N/A	N/A	>75
		Sentiment	N/A	N/A	>75
I	Acute and Inpatient	Theme	N/A	N/A	>75
		Sentiment	N/A	N/A	>75

N/A; Not applicable given that the final four organisations reached thematic saturation following the coding efforts of the initial five organisations and therefore achieved an overall accuracy rate of 75% or higher.

Three Trusts noted that sentiment prediction related to certain words were misclassified. The study team reviewed the performance and identified words that were culpable for this misclassification, e.g., “Nothing”, “Everything”, and “Ok”. These Trusts were advised to manually code a subset of 500 comments with equally distributed sentiments, especially focusing on the misclassified words, along with comments that represented an equal or almost equal number of comments per themes and/or sentiments to obtain a balanced dataset. To determine the accuracy of the algorithm, a pipeline was developed and refined to standardise the entire process within all the organisations as depicted in Figure 2.



© Imperial College Healthcare NHS Trust 2024. All rights reserved. For licensing enquiries contact nhsinfo@imperial.ac.uk

Figure 2. A pipeline created to develop and standardise overall accuracy calculation before final deployment.

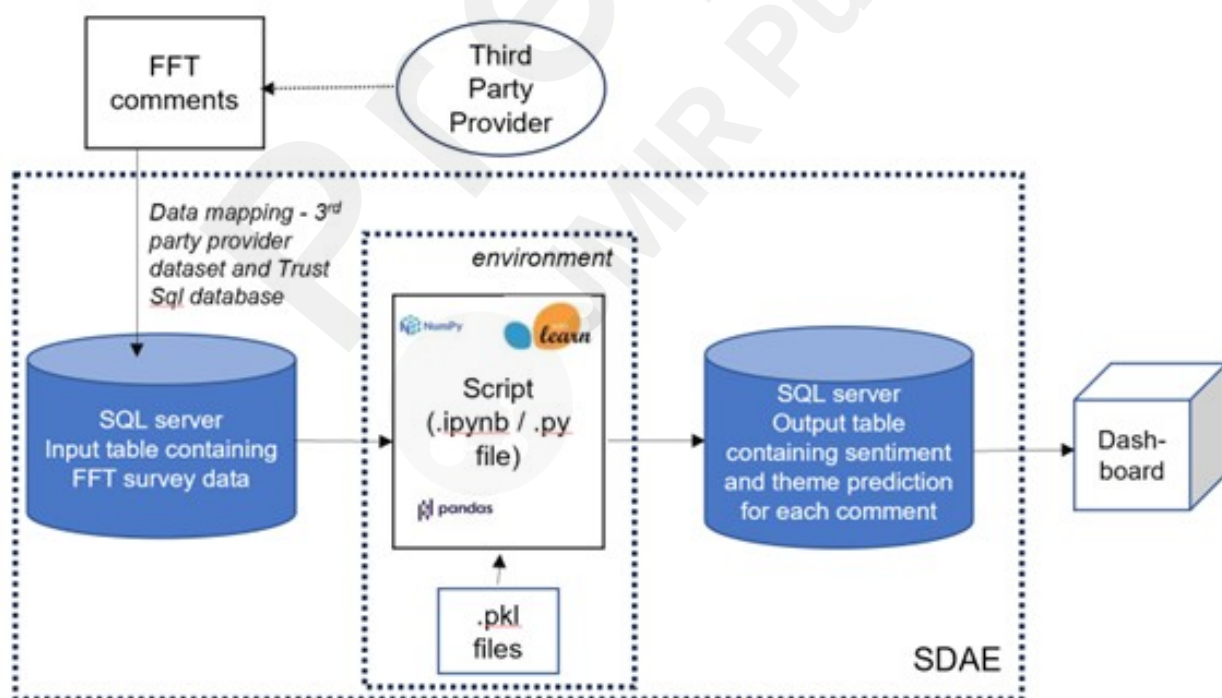
The overall accuracy for theme and sentiment significantly improved after using a revised coded template. As Trusts were added to the study sequentially, one by one, as part of the recruitment process, the BoW were also applied sequentially to the algorithm. Thematic saturation was achieved after analysing and collating comments from a total of five Trusts. This was observed when the algorithm was tested in four of the remaining Trusts without the need to implement their respective coding output and achieving an overall accuracy greater than or equal to 75%.

The final four organisations achieved an overall accuracy rate of 75% or higher, having reached thematic saturation following the initial coding efforts of the initial five organisations. As a result, as seen in Table 1 four versions of the model were deployed as new, labelled comments were collated from the different Trusts that joined the programme and deployed the algorithm. This iterative learning was done by adding randomly scrambled new unique words that the algorithm encountered in the BoW within its code. Each time, this word repository was completely randomised, and each word was treated individually by

removing any linkage. As a result, it was not possible to know from which Trust the word came from or if it was part of a positive or negative comment, with the study team holding the ground truth.

Adaptations and running without errors

The algorithm script was adapted according to agreements and requirements of the individual organisation. The solution tested in Jupyter Notebook for easier interaction with the code and checked for errors. The two most common errors encountered were “FileNotFoundError” which was a result of not specifying the correct path to a .pkl file, “Invalid column name” as the code had not been adapted appropriately with the column names from the input table on the SQL server. Once errors were corrected the final deployment consisted of the output with predicted themes and sentiment exported to a SQL database for integrating into dashboard visualisation. This pipeline (Figure 3) was standardised after refinement with all participating organisations. Each Trust automated their dataflow based on the requirements defined by the frequency and volume of comments that the pipeline should be updated with.



© Imperial College Healthcare NHS Trust 2024. All rights reserved. For licensing enquiries contact nhsinfo@imperial.ac.uk

Figure 3. Final pipeline demonstrating steps required to deploy the final algorithm with FFT comments as input, prediction of themes and sentiment using Pickle files generated from summative coding outputted to SQL database for use in dashboard visualisation.

For this purpose, user friendly interfaces were designed and developed using Flask Python library (29) for the backed and HTML, Css and Bootstrap for the frontend. Functionalities of these interfaces included analysing the accuracy of the algorithm by uploading a labelled dataset and re-training the algorithm with more coded data.

Algorithm refinement

Three hospitals noted that after manual inspection of the sentiment prediction some FFT comments were being classified with an incorrect sentiment. We reviewed the performance of the algorithm and identified words that were making the algorithm perform incorrectly (e.g. "Nothing", "Everything", "Ok"). It was suggested that these hospitals should manually code a subset of 500 comments with equally distributed sentiments, especially focusing on these words, along with comments that represented an equal or almost equal number of comments per themes and/or sentiments to obtain a balanced dataset. A code developed by the study team was provided to ensure the equal distribution of both themes and sentiments in the dataset.

Multi-tagging

Potential refinements to the algorithm, particularly regarding the need to break long comments into sentences for Trusts encountering lengthier narratives, such as those in mental health and community settings were highlighted and a multi-tagging method was devised. The multi-tagging approach involved assigning multiple themes from the NHS Patient Experience Framework to a single text document based on its content. Unlike traditional single-label classification, where the FFT free-text is assigned to only one label, multi-tagging recognises that comments may encompass various themes. This approach enables a more nuanced representation of the content's complexity and captures multiple dimensions of meaning within the same document. The punctuation marks, words and number of characters used to split the original sentence, can be customised based on different requirements using Spacy ($\geq 3.7.2$) and its English language model. As a default,

sentences are split after a maximum of 75 characters, and each of these is analysed individually (Figure 4). This can be then customised according to the requirements of the Trust to define how the base text should be split (e.g. by punctuation marks or specific words such as 'however'). Based on this level of customisability, other Trusts also implemented the solution.

Original Comment ID	Sentence	Theme	Sentiment
1	Amazing care from our general paediatrician who is going above and beyond to help coordinate multiple professionals in the care of our child. Seriously! I am so grateful for her expertise and wisdom and feel extremely blessed that our child is under her care. Grateful thanks. She has been so supportive of us as parents too.	1	1

↓ **Final Split Sentences**

Original Comment ID	Sentence	Theme	Sentiment
1	Amazing care from our general paediatrician who is going above and beyond to help coordinate multiple professionals in the care of our child.	1	1
1	Seriously! I am so grateful for her expertise and wisdom and feel extremely blessed that our child is under her care. Grateful thanks. She has been so supportive of us as parents too.	5	1

Figure 4: Example of how the sentencing approach works. The original comment is split into two sub-comments, each classified according to a theme and a sentiment. The original comment identifier is retained to preserve the metadata of the comment.

Redaction of personal and identifiable information

Considerable variation was also observed in how Trusts approached the redaction process, ranging from automated processes with subsequent manual checks to entirely manual methods. To address this inconsistency, a redaction algorithm capable of automatically detecting and masking personally identifiable information was devised. A redaction algorithm was developed using NLTK, Regex and Presidio Python libraries to automatically detect and mask the personally identifiable information. The algorithm can be customised according to the requirements of the Trust to define how and what data needs to be redacted.

Community of Practice

During the quarterly meetings, there was a high level of engagement from at least 80% of staff members representing the nine Trusts. This level of participation indicated strong interest and commitment to the community of practice's objectives. Additionally, feedback from participants revealed a positive perception of the community of practice, with at least 90% reporting a sense of belonging and support within the community. This sense of camaraderie and mutual support contributed to the effectiveness of the community of practice in addressing algorithm deployment challenges and driving improvements based on patient feedback. We addressed maintenance of the algorithm through the community of practice by creating a standardised software that enables each hospital to check the accuracy of the algorithm and make changes as needed. Figure 5 shows the interfaces developed.

Figure 5. Interfaces developed by the Community of Practice to assist with regular algorithm maintenance.

Discussion

Using a test-and-iterate approach, this study deployed a text analytics algorithm to predict theme and sentiment from FFT free text data in nine NHS organisations in England with different care settings, geographical location and demographics. Furthermore, the study highlighted the variation in technical capacity and resources among Trusts, introduced and developed a framework and standardised data processing pipeline capable of supporting FFT text analytics bespoke to the organisation's needs. The key contributions of this study are: (1) the construction of technical infrastructure to support and run the algorithm effectively and being able to adapt the script for individual organisations; (2) the creation of master coding template with free-text data from different healthcare organisations to support algorithm iteration and improve overall accuracy for theme and sentiment; (3) the creation of pipeline to standardise the method of incorporating the algorithm for processing free-text data; and (4) foster collaborative non-siloed working that encourages local ownership and share-learning.

Addressing data management and standardisation through information technology infrastructure

Data collection and analysis is central to this project. Achieving this required a robust data infrastructure with different components that collaboratively created an effective framework for research data management. This framework enhanced data accessibility by presenting information in a consistent, standardised, predictable and accessible format while also strengthening data reliability through the automation of manual processes.[22] However, ensuring high-quality data demands a diverse set of information management skills, including data acquisition, storage, organisation, retrieval, and long-term maintenance.[23] This project offered a valuable opportunity for Trusts to gain expertise in this area by providing guidelines to streamline processes, define roles, and establish a resilient information technology infrastructure.

Creating a template with additional organisational specific free-text data

Each healthcare setting possesses unique attributes, patient demographics, and linguistic nuances as demonstrated above. Algorithm testing was required to allow for customization

and tailoring of the algorithm to accommodate these contextual variations. As part of a supervised learning approach, re-coding was required to re-establish ground truth and new words were incorporated into an expanded BoW to improve performance across various care settings. Literature suggests that the larger the training sets used, the higher the accuracy of the algorithms at identifying similar comments within the broader dataset, but trade-offs with time and human coding are necessary to ensure the method is resource-efficient.[10] However, the nature of patient experience vocabulary, due to the domain of patient feedback from free-text is fixed in its nature, making it attractive data for supervised learning.[24] Therefore, it is possible to anticipate the meaning of various phrases and automatically classify the comments.[25] Thematic saturation was reached with five organisations and no further coding was necessary. Just as the domain is fixed, the perspective of a patient feedback document is also fixed; there is limited vocabulary that is useful for commenting about health services.[24] Rastegar-Mojarad et al. [26] also observed that a small (25%) vocabulary set covered a majority (92%) of the content of their patients' comments, consistent with a study exploring consumer health vocabulary used by consumers[27]. This suggests that patients use specific vocabulary when expressing their experience within free-text comments.

In settings that were relatively fixed, i.e. adult inpatient, the classifier trained strong reusable models that was evident from satisfactory accuracy. However, the accuracy of the model was lower in organisations delivering different services such as paediatrics and mental health. This was to be expected given the language and type of words used in different care settings would vary from those found in an adult inpatient setting. The customization ensured that the algorithm aligns with the specific language and themes prevalent in the patient feedback of a particular healthcare organisation. The accuracy for theme improved in all organisations demonstrating that whilst the algorithm may not initially be transferable, with some retraining and adaptations it can easily be reused in different service settings. As part of the algorithm development process, cross-validation was applied to ensure the robustness and reliability of the model within the dataset used. In addition to this, we conducted external validation by incorporating patient comments from multiple Trusts, ensuring that the model was exposed to a diverse range of data prior to declaring thematic saturation. This strategy allowed us to evaluate the model's performance beyond a single dataset and ensured that it was not overfitting to any Trust or setting. By including data from

various Trusts, we created a working external validation set, which will serve as the foundation for assessing the model's generalizability in future implementations.

Furthermore, 'respect for patient centred values' had the highest count in all organisations. This finding is similar to the literature around patient experience reporting in healthcare.[28-32] The reasons for this are twofold, firstly, primarily due to the large number of interchangeable words found within that theme, i.e., 'Respect for patient-centred values, preferences, and expressed needs (cultural issues, dignity, privacy and independence, awareness of quality-of-life issues and shared decision making)'. Secondly, the overall domain of patient feedback is the healthcare system, and this study revealed that the content of reviews tends to focus on a small collection of aspects associated with this, as demonstrated by the themes used for text classification in the studies.

Through community of practice participation, multi-tagging was used to assign multiple relevant tags or categories to individual sentences within a single FFT free-text comment. Instead of assigning a single label to the entire document, this technique involved breaking down text into sentences and assigning appropriate themes and sentiments to each based on its content, thereby enhancing customisability within mental health and community care settings. Furthermore, the community of practice assisted in developing a redaction solution using Python libraries which can be then customised according to the requirements of the Trust to define how and what data needs to be redacted. These solutions can subsequently be adopted by other Trusts as an improvement to their existing processes.

Creating a pipeline to standardise the algorithm deployment approach

As part of the model testing and to make the model more widely re-usable, we created a data pipeline, which infuses and integrates BoW from different Trusts to predict theme and sentiment at an accuracy greater than 75% when used in different Trusts. This pipeline serves as a blueprint to stakeholders and facilitates understanding of all pieces of the pipeline within the team members and standardisation when deploying the model. Compliance with regulations and governance is supported, and integration with DevOps practices enhances machine learning model integration into the software development lifecycle.

Generating cultural change to prioritise patient experience centred care

The establishment of CoP can play a pivotal role in establishing a multi-disciplinary team of stakeholders from IT, business intelligence, patient experience, and QI within each Trust. Furthermore, the CoP can provide the necessary guidance to use the algorithm and assist in developing a dashboard for quality improvement projects,[14] and sustained enthusiasm in the use of data and data analytics to improve the quality of care for patients and carers. [33] This cultural transformation will result in increased organisation agility and responsiveness to patient needs, driving positive changes in care delivery practices.[34]

Limitations

The Pickle files used in this study are specific to the version of Python and the libraries used to create them. Attempting to load a pickle file created with an older version of Python or with different library versions, may encounter compatibility issues. Newer versions of Python may have introduced changes to the pickle module, or the libraries being pickled, which can result in errors when loading the file. To mitigate challenges with loading an outdated pickle file, updating the code that creates the pickle file or alternative serialization formats that are more version-independent can be considered. The model can be containerized using Docker with a requirements file containing all the required libraries and Docker specifying the environment setup and start-up operations. The Docker Container can then be deployed to application hosting services, whether on cloud or on-premises servers.

The limited vocabulary range is a significant issue faced by the BoW model. For example, if the model encounters an unfamiliar or rare but informative word that it has not yet seen during training, BoW model will tend to ignore it. To mitigate this limitation, we trained the model on five out of nine Trusts to ensure that the model has gathered a substantial BoW representation, further augmented by thematic saturation. Significant change to patient demographics or significant changes such as COVID-19 pandemic could reduce the performance due to unfamiliar words. Therefore, regular updates in the model would be required to keep it up to date. For this purpose, user friendly interfaces can be designed and developed using Flask Python library for the backend and HTML, Css and Bootstrap for the frontend. Functionalities of these interfaces included analysing the accuracy of the algorithm by uploading a labelled dataset and re-training the algorithm with more coded

data. The interfaces allow organisations to periodically check accuracy as part of a responsible artificial intelligence strategy.

Given the limited analytical expertise and computational resources in many UK hospitals, complex models like Transformers and Recurrent Neural Networks (RNNs) would be difficult to implement effectively. The SVM model was chosen for its simplicity, efficiency, and low resource demands, making it ideal for integrating into existing workflows. The focus was on maximising the value of patient feedback data to improve care quality, without overwhelming the hospital's IT infrastructure. We acknowledge that alternative lightweight machine learning methods, such as Decision Forests and XGBoost, offer similar efficiency and computational efficiency without the significant computational burdens often associated with deep learning models. Furthermore, modern deployment strategies, such as using pre-trained architectures and lightweight deep learning models, can help mitigate the resource challenges typically associated with training deep learning models like RNNs. As the hospital's analytics capacity grows, more sophisticated models can be explored, but SVM provided a practical starting point for immediate, actionable insights.

The authors set out to illustrate the process of developing an NLP pipeline that originated in a single institution and has been iteratively refined for broader implementation. A strong technical knowledge base was essential to address variations in technical capabilities and IT infrastructure, as well as to introduce and enhance analytical skills that had been underutilized in some hospital Trusts. However, to successfully implement the intervention within complex healthcare systems, it is crucial to understand how the intervention interacts with and influences various system components to achieve the desired outcomes. We are currently collaborating with participating Trusts to establish a data flow that automates the integration of NLP outputs into a dashboard. Additionally, we are making algorithm improvements based on local priorities and needs to foster local ownership of the process.

Responsible Artificial Intelligence

As artificial Intelligence (AI) continues to advance, becoming increasingly integrated into healthcare systems, ethical considerations and technical challenges become increasingly intricate. For this reason, a key aspect of the project included providing tools and best practices to ensure that AI is being implemented responsibly. Firstly, the developed user-

friendly interfaces, promote the ongoing monitoring of the accuracy of the algorithm, ensuring fairness and accountability of the version of the model that was deployed. Secondly, the establishment of CoP will play a critical role in fostering a culture of continuous improvement and supporting the transformation towards greater organizational agility but also cultivates an environment where professionals can delve into the nuanced aspects of responsible AI. This transformation will enhance the responsiveness of healthcare systems to patient needs and drive positive changes in care delivery practices. Furthermore, by building local expertise and ownership, we aim to create a sustainable framework for ongoing use of data and analytics to improve the quality of care. This contributes to the development and refinement of ethical guidelines, playing a crucial role in shaping the responsible AI landscape.

Future directions

The study group are currently conducting a complex, multi-component implementation designed to enhance operational efficiencies within the NHS by empowering frontline staff with visible and actionable insights derived from patient feedback tool, which aligns with the NHS FFT re-development policy recommendations. This intervention aims to drive sustainable impact and foster a culture of patient-centred working by facilitating the identification and implementation of improvement opportunities. The CoP will play a vital role in achieving the broader goal of embedding a patient-centred working culture across the NHS. They will help sustain momentum for using patient feedback and data analytics to drive continuous improvements in care. This multi-faceted approach ensures that the intervention not only has immediate operational benefits but also creates a lasting cultural shift towards patient-centred care.

Models like BERT have the potential to offer deeper insights by capturing more complex linguistic patterns, which could be particularly valuable for understanding the subtleties in patient comments, especially in sensitive areas such as mental health. While we considered exploring these advanced models, we opted for a more straightforward approach during this phase of implementation. This decision was driven by the need for a method that is both practical and easily scalable across diverse healthcare settings. However, future research with a more experienced group, enriched by the extended knowledge gained through the CoP, could involve comparing transformer-based models like

BERT with the current approach to evaluate whether they provide significant improvements in accuracy and interpretability. Moreover, future work will explore the impact of different preprocessing strategies, such as retaining certain negations, expanding contractions or adjusting tokenization rules (including experimenting with different n-gram configurations) based on contextual understanding. This would allow for a more nuanced approach, potentially improving the accuracy of sentiment classification, especially in cases where subtle cues are essential.

To enhance the interpretability of the model's predictions, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be incorporated into the NLP pipeline. SHAP can explain that the words "excellent" and "care" contributed significantly to a positive prediction, while "waited" and "long" contributed to a negative prediction. This helps the healthcare team understand the key drivers behind feedback, allowing them to make targeted improvements. For a piece of patient feedback stating, "The doctor was kind, but the wait was too long," LIME might perturb the sentence by removing words like "kind" or "wait" to see how the model's prediction changes. If the model changes from positive to negative after removing the word "wait," LIME would indicate that the term "wait" is highly important for determining the sentiment of the feedback. This could be an important step for Trusts in improving transparency which builds trust and helps the healthcare team make better-informed decisions when observing the dashboard.

Conclusion

This study marks an advancement in harnessing free-text FFT data to gain valuable insights in healthcare settings through the creation of a robust supervised learning text analytics algorithm. The observed disparities in certain care settings were expected, considering the inherent differences in lexicon and terminology compared to the adult inpatient care setting where the algorithm was initially developed. Addressing these challenges involved additional coding and thorough testing under diverse scenarios. Through this iterative process, the accuracy and reliability of the algorithm were established to be robust and easy to use, fostering inter- and intra-organisational collaboration, and promoting shared learning within the healthcare domain.

Funding

This work was funded by the Health Foundation and supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre and the NIHR Imperial Patient Safety Research Centre. The research was enabled by the Imperial Clinical Analytics Research and Evaluation (iCARE) environment and used the iCARE team and data resources. Development of the IT scoping tool and coding pack was funded by NHS England Insight and Feedback Team. The study funder(s) did not play a role in study design; in the collection, analysis, and interpretation of data; in writing of the report; and in the decision to submit the article for publication. In addition, researchers were independent from funders, and all authors had full access to all the data included in this study and can take responsibility for the integrity of the data and accuracy of the data analysis.

Acknowledgments

The authors would like also to like to acknowledge the contributions of all individuals from

the NHS Trusts involved in the Scale, Spread and Embed Programme:

Victor Banda, William Bryant, Suzanne Collin, Zeina Dahan, Paul Gough, Mark Harris, Taraben Kapadia, Pavithra Rajendran, Jonny Sheldon, Caitlin McGovern, Marcus Beck, Ahmed Sharif, Mandy Dawley, Catherine Hunter, Brad Walker, Mark Ward, Luca Mercuri, Rohini Subramaniam, Ben Glampson, Amy Bond, Rosie Frost, Paul Morris, Emmanuel Okafor, Andrew Walsh, Dean Wild, Lesley Whittaker, Quinta Ashcroft, Vishnu Chandrabalan, Stuart Clough, Rebecca Hargreaves, Joan Ravencroft, Stephanie Iaconianni, John Howles, Karen Greenwood, Ben Morris, Lisa Bonifacio, Sarah Eastwood, Nathan Harding, Paul Hooker, Jacek Pawlowski, Julie Cheney, Ana Maria Aldana Gelves, Prajwal Khairnar, Trudy Taylor, Sarah Curtis, Simon Godley, Chris Lewis, Amy Marchant, Mary Mason, Oluwatoni Ojelade, Stephen Onamade, Vijay Rabadia, Chloe Roebuck, Francesca Roffey, Saad Saleem, Neil Springham, Jon Blacker, Edmund Gall, Matthew Stammer, Michael George, Laura White

In addition, we are very grateful to members of the Programme Steering Board for their oversight and constructive feedback during the Programme roll-out.

Appendix



References

1. Flott KM, Graham C, Darzi A, Mayer E. Can we use patient-reported feedback to drive change? The challenges of using patient-reported feedback and how they might be addressed. *BMJ Qual Saf.* 2017;26(6):502-7. Epub 20160620. doi: 10.1136/bmjqs-2016-005223. PubMed PMID: 27325796.
2. England N. The Friends and Family Test. NHS England, 2014 Contract No.: Publication Gateway Ref No. 01787.
3. Gleeson H, Calderon A, Swami V, Deighton J, Wolpert M, Edbrooke-Childs J. Systematic review of approaches to using patient experience data for quality improvement in healthcare settings. *BMJ Open.* 2016;6(8):e011907. Epub 2016/08/18. doi: 10.1136/bmjopen-2016-011907. PubMed PMID: 27531733; PubMed Central PMCID: PMC5013495.
4. Usman M, Mujahid M, Rustam F, Flores E, Vidal Mazon JL, Diez IT, Ashraf I. Analyzing patients satisfaction level for medical services using twitter data. *PeerJ Comput Sci.* 2024;10:e1697. Epub 20240109. doi: 10.7717/peerj-cs.1697. PubMed PMID: 38259896; PubMed Central PMCID: PMC5013495.
5. Using the Friends and Family Test to improve patient experience. NHS England and NHS Improvement, 2019.
6. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf.* 2013;22(3):251-5. Epub 2013/01/26. doi: 10.1136/bmjqs-2012-001527. PubMed PMID: 23349387.
7. Khanbhai M, Warren L, Symons J, Flott K, Harrison-White S, Manton D, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. *Int J Med Inform.* 2022;157:104642. Epub 20211111. doi: 10.1016/j.ijmedinf.2021.104642. PubMed PMID: 34781167.
8. Devlin J, Chang, M., Lee, K., Toutanova, K., editor BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Conference of the North American Chapter of the Association for Computational Linguistics; 2019 June 2-7; 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis2019.
9. Haleem A, Javaid, M., Singh, R. Exploring the competence of ChatGPT for customer and patient service management. *Intelligent Pharmacy.* 2024;2(3):392-414.
10. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform.* 2021;28(1). doi: 10.1136/bmjhci-2020-100262. PubMed PMID: 33653690; PubMed Central PMCID: PMC5013495.
11. Sakai H, Lam, S., Mikaeili, M., Bosire, J., Jovin, F. Large Language Models for Patient Comments Multi-Label Classification. *arXiv.*
12. Osvath M, Yang G Z., Kosa, K. Analyzing Narrative of Patient Experiences: A BERT Topic Modelling Approach. *Acta Polytechnica Hungarica.* 2023;20(7):153 - 71.
13. Strubell E, Ganesh, A., McCallum, A., editor Energy and Policy Considerations for Deep Learning in NLP2019: Association for Computational Linguistics.
14. Khanbhai M, Symons J, Flott K, Harrison-White S, Spofforth J, Klaber R, et al. Enriching the Value of Patient Experience Feedback: Web-Based Dashboard Development Using Co-design and Heuristic Evaluation. *JMIR Hum Factors.* 2022;9(1):e27887. Epub 20220203. doi: 10.2196/27887. PubMed PMID: 35113022; PubMed Central PMCID: PMC5013495.
15. Enston C. The power of the written word2021. Available from: <https://www.england.nhs.uk/>

blog/the-power-of-the-written-word/.

16. Pedregosa F, Varoquax, G., Gamfort, A., et al. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*. 2011;12:2825 - 30.
17. Board NNQ. NHS Patient Experience Framework. Department of Health, 2011.
18. Sparck Jones K. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*. 1972;28(1):11-21. doi: 10.1108/eb026526.
19. Pang B, Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. 2008;2:1-135.
20. Duarte JM, Berton L. A review of semi-supervised learning for text classification. *Artif Intell Rev*. 2023;1-69. Epub 20230131. doi: 10.1007/s10462-023-10393-8. PubMed PMID: 36743267; PubMed Central PMCID: PMCPMC9887265.
21. Denecke K, Reichenpfader D. Sentiment analysis of clinical narratives: A scoping review. *J Biomed Inform*. 2023;140:104336. Epub 20230322. doi: 10.1016/j.jbi.2023.104336. PubMed PMID: 36958461.
22. Towbin AJ. Collecting Data to Facilitate Change. *J Am Coll Radiol*. 2019;16(9 Pt B):1248-53. doi: 10.1016/j.jacr.2019.05.032. PubMed PMID: 31492402.
23. Neumann J. FAIR Data Infrastructure. *Adv Biochem Eng Biotechnol*. 2022;182:195-207. doi: 10.1007/10_2021_193. PubMed PMID: 35091812.
24. LaVela S, Gallan, AS. Evaluation and measurement of patient experience. *Patient Experience Journal*. 2014;1(1):28-36.
25. Garrisson M WJ. The role of volunteer in improving patient experience. Dallas TX: The Beryl Institute 2016. Available from: <http://www.theberylinstitute.org/news/284517/The-Role-of-the-Volunteer-in-improving-Patient-Experience-Explored-by-The-Beryl-Institute-.htm>.
26. Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S. Collecting and Analyzing Patient Experiences of Health Care From Social Media. *JMIR Res Protoc*. 2015;4(3):e78. Epub 2015/07/04. doi: 10.2196/resprot.3433. PubMed PMID: 26137885; PubMed Central PMCID: PMCPMC4526973.
27. Indovina K, Keniston A, Reid M, Sachs K, Zheng C, Tong A, et al. Real-time patient experience surveys of hospitalized medical patients. *J Hosp Med*. 2016;11(4):251-6. doi: 10.1002/jhm.2533. PubMed PMID: 26777721.
28. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf*. 2016;25(6):404-13. Epub 2015/10/16. doi: 10.1136/bmjqs-2015-004309. PubMed PMID: 26464518; PubMed Central PMCID: PMCPMC4878682.
29. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, Millett C. Tweets about hospital quality: a mixed methods study. *BMJ Qual Saf*. 2014;23(10):838-46. Epub 2014/04/22. doi: 10.1136/bmjqs-2014-002875. PubMed PMID: 24748372; PubMed Central PMCID: PMCPMC4174012.
30. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Qual Manag Health Care*. 2012;21(1):9-19. Epub 2011/12/31. doi: 10.1097/QMH.0b013e3182417fc4. PubMed PMID: 22207014.
31. Nawab K, Ramsey G, Schreiber R. Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback. *Appl Clin Inform*. 2020;11(2):242-52. Epub 2020/04/03. doi: 10.1055/s-0040-1708049. PubMed PMID: 32236917; PubMed Central PMCID: PMCPMC7113005.
32. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal

cancer care. *BMJ Qual Saf.* 2016;25(8):604-14. Epub 2015/10/30. doi: 10.1136/bmjqs-2015-004063. PubMed PMID: 26512131.

33. Dowding D, Randell R, Gardner P, Fitzpatrick G, Dykes P, Favela J, et al. Dashboards for improving patient care: review of the literature. *Int J Med Inform.* 2015;84(2):87-100. Epub 2014/12/03. doi: 10.1016/j.ijmedinf.2014.10.001. PubMed PMID: 25453274.

34. P P. What gets measured gets done. Or does it? 2014. Available from: <https://www.health.org.uk/blogs/what-gets-measured-gets-done-or-does-it>.



Supplementary Files

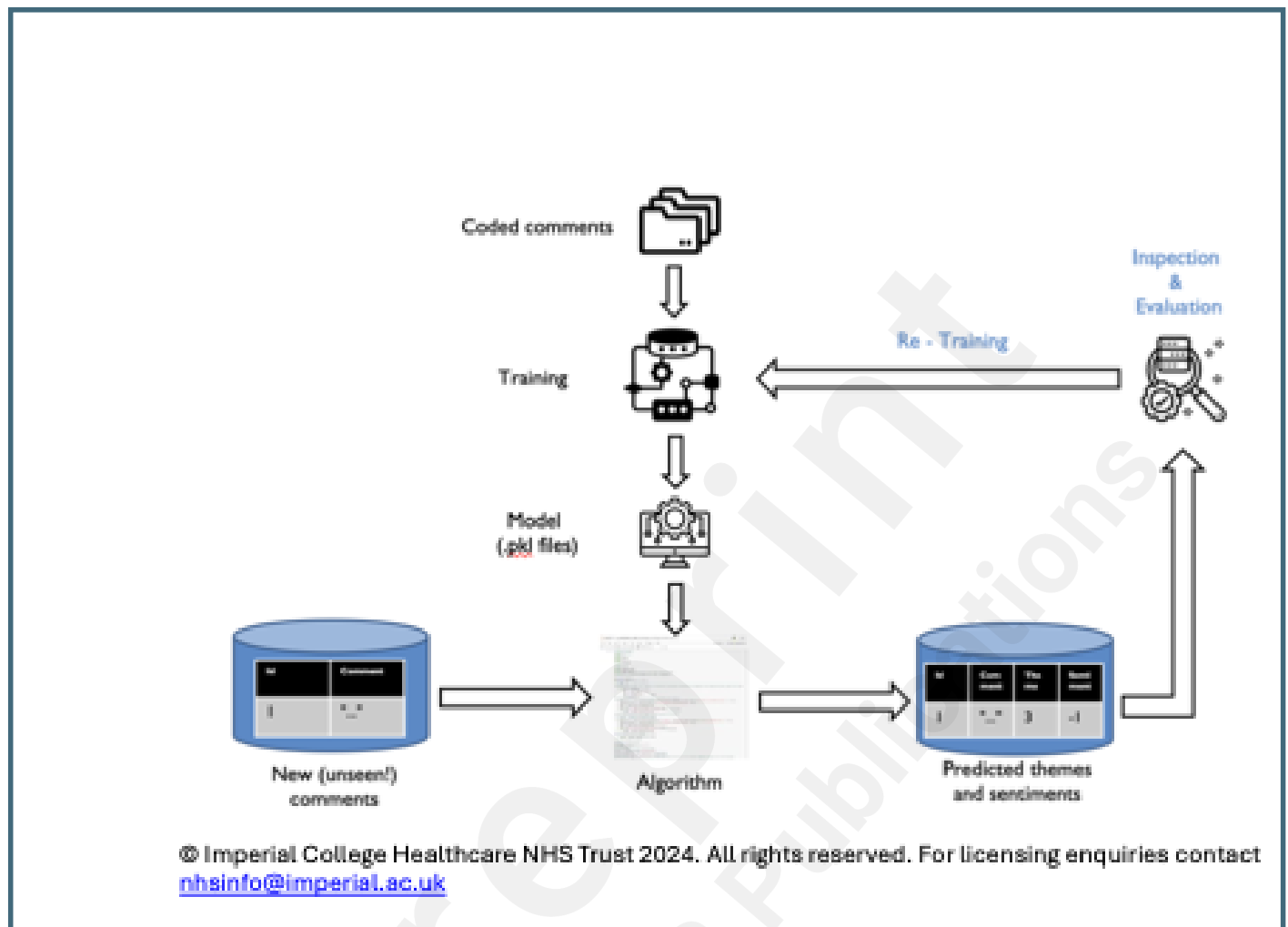
Figures

An overview of the geographical distribution of the nine recruited Trusts.

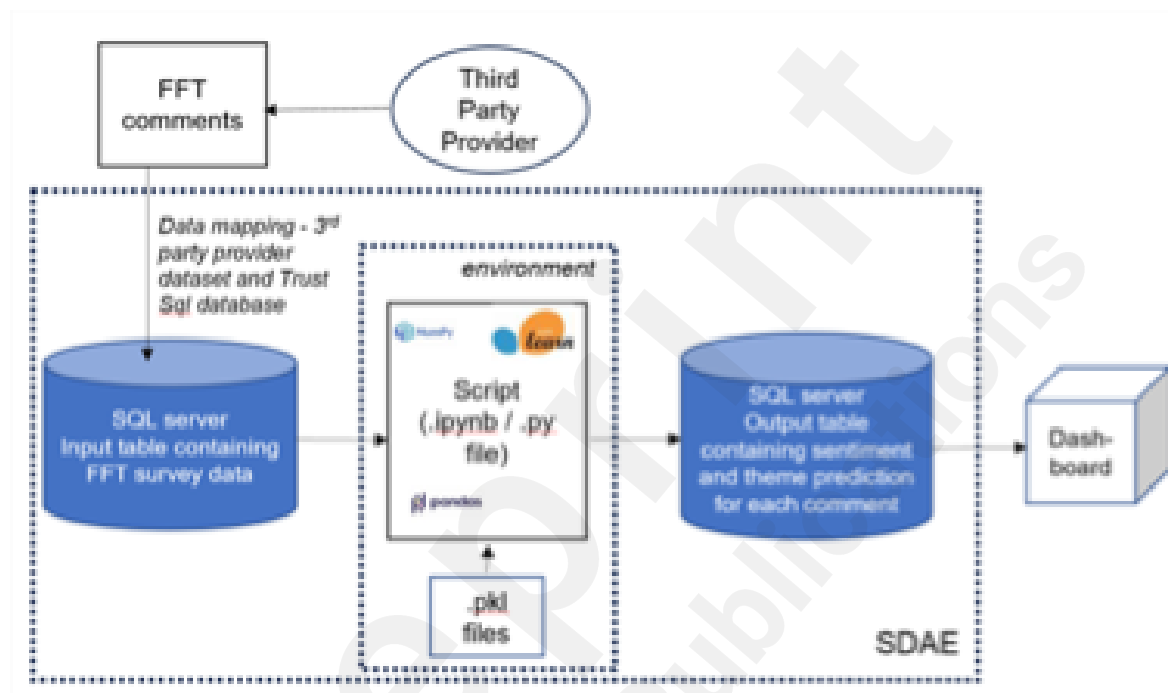


© Imperial College Healthcare NHS Trust 2024. All rights reserved. For licensing enquiries contact nhsinfo@imperial.ac.uk

A pipeline created to develop and standardise overall accuracy calculation before final deployment.



Final pipeline demonstrating steps required to deploy the final algorithm with FFT comments as input, prediction of themes and sentiment using Pickle files generated from summative coding outputted to SQL database for use in dashboard visualisation.



© Imperial College Healthcare NHS Trust 2024. All rights reserved. For licensing enquiries contact nhsinfo@imperial.ac.uk

An example of how the sentencing approach works. The original comment is split into two sub-comments, each classified according to a theme and a sentiment. The original comment identifier is retained to preserve the metadata of the comment.

Original Comment ID	Sentence	Theme	Sentiment
1	Amazing care from our general paediatrician who is going above and beyond to help coordinate multiple professionals in the care of our child. Seriously! I am so grateful for her expertise and wisdom and feel extremely blessed that our child is under her care. Grateful thanks. She has been so supportive of us as parents too.	1	1

 **Final Split Sentences**

Original Comment ID	Sentence	Theme	Sentiment
1	Amazing care from our general paediatrician who is going above and beyond to help coordinate multiple professionals in the care of our child.	1	1
1	Seriously! I am so grateful for her expertise and wisdom and feel extremely blessed that our child is under her care. Grateful thanks. She has been so supportive of us as parents too.	5	1

Interfaces developed by the Community of Practice to assist with regular algorithm maintenance.

FFT Algo Home Analyse Comments Get Accuracy

Welcome to the FFT Algorithm Testing Site

This interface will let you add an input .csv file with Patient feedback comments and get an associated theme and sentiment. The .csv file should have a column with an id and a column with the comments. Once the algorithm finishes working, an output file will be saved at the current path, with two extra columns for the assigned theme and sentiment.

[Process your FFT File](#)



Imperial College Healthcare NHS Trust aoua NIHR | National Institute for Health Research The Wellcome Foundation Imperial College London

Obtaining accuracy

This interface will let you add an input .csv file with Patient feedback comments and get an associated theme and sentiment. The .csv file should have a column with an id and a column with the comments. Once the algorithm finishes working, an output file will be saved at the current path, with two extra columns for the assigned theme and sentiment.

[Process your FFT File](#)

Analyse Comments

This interface will let you add an input .csv file with Patient feedback comments and get an associated theme and sentiment. The .csv file should have a column with an id and a column with the comments. Once the algorithm finishes working, an output file will be saved at the current path, with two extra columns for the assigned theme and sentiment.

[Process your FFT File](#)

Multimedia Appendixes

Untitled.

URL: <http://asset.jmir.pub/assets/b4778e83e620fa3b22b551567413d92b.pdf>

