Text mining FFT data- final report

Introduction

NHS provider organisations have collected an incredibly large amount of feedback from patients as part of the FFT programme. Many organisations collect a lot of text data in this way but lack the staff time to analyse it all systematically. In a lot of organisations the data will be read by service managers and others who are interested in patient experience but it is not systematically codified in a way that would make it possible to count, filter, and summarise the data with respect to its subject matter.

Machine learning can be used to read text and apply labels to it automatically (after appropriate training with a labelled dataset) and this project sought to produce an algorithm which is capable of ingesting text data and outputting two tags for each piece of text- its theme (environment/ facilities, staff, access to services…) and its *criticality* (how positive or negative a piece of feedback is- from "This service is appalling" through "The experience met my expectations" and on to "Everyone I met was a credit to their organisation". For a full list of theme and criticality labels please see Appendix 2.

This project aimed to produce two types of output. Firstly, a machine learning model written in open licensed code which can be used in this project and freely elsewhere to others in order to classify FFT text feedback according to its theme and criticality. Secondly, a dashboard in order to allow users of the data to interact with the text, filtering and aggregating by theme and criticality and helping them to explore the themes in the data and the actual text feedback associated with each theme. Patient experience data is highly subjective and emotive and this work is intended to support and streamline human interpretation of this data, and not to replace it. The methods developed as part of this project should help the users of this data to find data that is of interest to them, collecting feedback of different types together and allowing it to be read and understood by staff, carers, and patients who are interested in this data.

Informal review of existing solutions in the area of text mining patient feedback showed that there are proprietary solutions available. These solutions, however, all require ongoing payments to be made by the organisations who are using them, and generally force providers to use the paid feedback collection and collation system provided by the vendor. The aim of this project was to produce free outputs that could moreover be decoupled from the methods by which patient feedback is collected, stored, and reported. This would allow groups across providers and users of patient feedback systems to make use of the outputs of the project so as to maximise the benefits for their particular use case. The work was released under the permissive open source MIT licence to ensure that it could be freely reused by any organisation which could make use of it. The text of the MIT licence used in the project can be seen here https://github.com/CDU-data-science-team/pxtextmining/blob/main/LICENSE. For more details on MIT and other licences see https://tldrlegal.com/license/mit-license. This project is compatible with guidance from NHSX that is in draft at the time of writing which states that "All new source code that we produce or commission should be open and reusable by default: such that anyone can freely access, use, modify, and share the relevant code for any purpose" (see https://github.com/nhsx/open-source-policy).

Text mining

As described in the introduction, this project has produced three distinct elements:

1. **Text classification model** https://github.com/CDU-data-science-team/pxtextmining**.** A ML model that is able to predict the themes and criticalities of unlabelled feedback text
2. **Text mining dashboard** https://github.com/CDU-data-science-team/experiencesdashboard. An interactive dashboard that presents the results of the text classification model, along with further insights derived from the feedback text, including sentiment analysis and other text mining methods
3. **Model summary dashboard** https://github.com/CDU-data-science-team/pxtextminingdashboard. A dashboard designed to allow non technical users to assess the accuracy of the model in use, including summary metrics as well as example of classified data to enable the user to make their own judgements about accuracy

Text classification model

Let's assume that we are teaching a child to distinguish between cats and elephants. We can show the child different photos of the two animals in order to help them understand what is unique to each animal. For example, elephants are grey, massive animals with big ears and a trunk. Cats are much smaller and furry, and have small ears.

Given this information, the child learns (or is "trained") to distinguish between cats and elephants and, when given photos of the two animals that they have never seen before, they are most likely going to be able to say which is of a cat and which is of an elephant. They may not always get it right (after all, they are still learning), but they probably will be right most of the time.

Text classification works in a similar way. First, instead of child we have a mathematical model. Second, what would be the pictures in the example above are now the tags/themes that we have assigned to each feedback text. For example, if the patient is complaining that the ward was too cold, then the text would be tagged as "Environment/ facilities". If they are praising the communication abilities of a nurse, then the text would be tagged as "Communication". Third, the content of the text itself plays the role of the animal characteristics mentioned earlier. For instance, it would not be surprising that much of the feedback about "Environment/ facilities" had in it words such as "cold", "hot", "comfortable", "air condition", "heating" or phrases such as "room too hot". Therefore, the model can "learn" (i.e. it is "trained") to make associations between words/phrases and tags/themes- just like the child learns to make associations between animal characteristics and photos. The trained model can be then given new, unlabelled, feedback text and predict its tag/theme with a certain degree of accuracy.

What is and is not possible

Despite the impressive ability of a model to predict tags for a given text, and the satisfactory performance of our own models, there are some limitations that need to be discussed.

Number of tags

A model cannot possibly predict with good accuracy as many tags as we like. We need to make sure that the number of tags is reasonably large to capture the diversity in the feedback comments, but also sufficiently small to help the model make meaningful predictions. For example, more than 15 tags may be too many, whereas less than six may be too few.

The coders currently use 44 themes which can be collapsed into the following nine higher-level themes that we used in our model: Access, Care received, Communication, Couldn't be improved, Dignity, Environment/ facilities, Miscellaneous, Staff, and Transition/coordination. For criticality, there are 11 values, from -5 to 5, including zero.

We built two ML models: one that predicts the nine themes, and one that predicts nine criticality values (from -4 to 4- see section “Lack of data”). Both models perform satisfactorily on the ~10k rows of data that were available at the time of writing.

Multiple tags for each feedback text

It is not surprising that in their feedback patients often talk about more than one issue, which could fall under one, two or more themes. Currently, our coders assign two themes (where applicable) to the text, the first one being the most “prevalent” in the text. Currently, the ML model is trained on the first theme for each feedback text. However, it is possible to train a model on both themes for each feedback text. Future work on this so-called “multi-label” modelling approach would further improve the performance of the model.

Lack of data

There are very few comments with a criticality of -5 or 5, rendering impossible the fitting of a ML model that performs well. We therefore replaced criticalities -5 and 5 with -4 and 4 respectively in the relevant text.

Assessing model performance

In order to assess the predictive performance of a model we use about 70% to train the ML model. This chunk of data is known as the *training dataset*. We then use the model to predict the tags for the remaining 30% of the data. This chunk of data is called the *test dataset*. Because we already know the tags in the test dataset, we can compare the actual and predicted tags to tell how well the model did in predicting them.

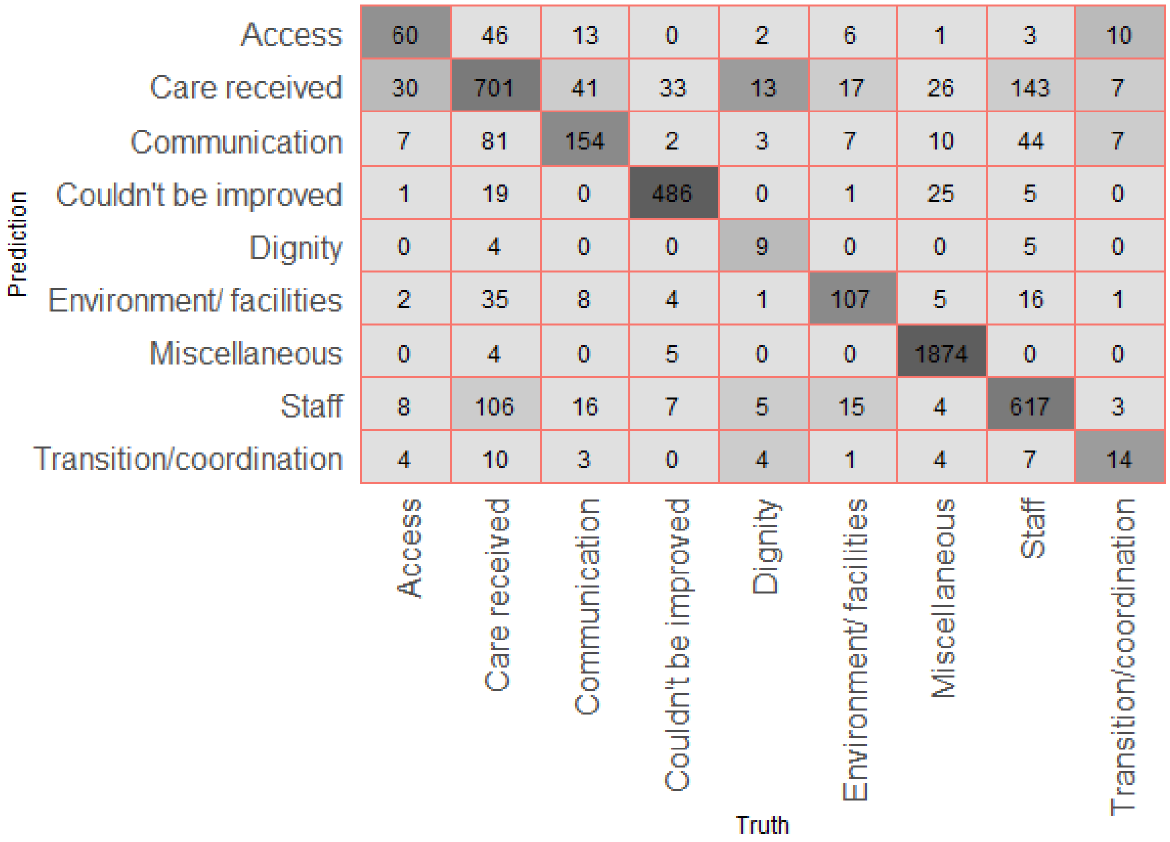
But how do we actually *measure* model performance? Naturally, we can just count the number of correct predictions in the test dataset and divide it by the number of records in that dataset to get an indicator of model *accuracy*. This is a standard, intuitive and easy-to-communicate way of quantifying model performance. It does come with a major drawback though: if a model is good at predicting only a few tags for which plenty of records are available, and is bad at predicting the rest of the tags, then the accuracy score will mask this. Going back to the cats *versus* elephants example, let’s assume that the child correctly identifies 950/1000 (95%) photos of elephants but only 325/500 (65%) photos of cats. In other words, they are great at identifying elephants but not that good at identifying cats. However, the overall accuracy is (950 + 325) / (1000 + 500) = 85%, which can mislead us in thinking that the overall result is a pretty good one.

It is important to mention that accuracy’s lack of ability to account for such imbalanced tag-specific performances is not necessarily a drawback- it very much depends on the end goal. If the purpose is to correctly predict the tags for as much of the feedback text as possible, regardless of the fact that prediction would be poor for some tags and great for others, then accuracy is an appropriate choice. On the other hand, if it is important to accurately predict as many tags as possible for *each* tag, then accuracy is clearly inappropriate. Following confirmation from our internal and external user groups and partners it was agreed that an alternative measure would be more appropriate than simple accuracy.

There are ways to assess the predictive performance of a model that do not suffer from the shortfalls of accuracy. Three are worth mentioning here: *class balance accuracy*, *confusion matrix* and *accuracy per class*. The first one, *class balance accuracy* (Mosley, 2013), is like an accuracy score, although corrected to account for such severe cases where the model does not predict well some or many of the tags. We used class balance accuracy to fit the model, i.e. we chose the model that had the highest class balance accuracy during training.

The second one, the *confusion matrix*, is a way to visually detect which tags the model has performed best/worst in predicting. The actual tags are on the rows and the predicted ones are on the columns (see figure). With a perfect model we would see zeros in all cells other than the diagonal. This would mean that the model predicts all tags correctly. In reality, a perfect prediction accuracy is rarely the case. Thus, in practice, with good model we would see most records on the diagonal, as this would mean that most records are predicted correctly.

In the figure, the numbers correspond to the number of records belonging to the tag on the rows that were predicted as the tag on the columns. The shades of grey translate these counts into proportions of the total number of records in the tag on the rows. With a good model, the darker shades of grey would be on the diagonal. The shades help spot where the most severe misclassifications may have occurred. They thus help see if the model is consistently confusing the text for a tag on the row as being about a different tag on the column. This is particularly important for assessing how "far off" the model's misclassifications are from reality. For instance, it seems like "Staff" is often misclassified as "Care received". This is not a severe misclassification error, since often the care received will depend on the staff delivering it. If, however, "Staff" was often misclassified as "Environment/ facilities", then this would be an indication that the model is performing poorly.



	Access	Care received	Communication	Couldn't be improved	Dignity	Environment/ facilities	Miscellaneous	Staff	Transition/coordination
Access	60	46	13	0	2	6	1	3	10
Care received	30	701	41	33	13	17	26	143	7
Communication	7	81	154	2	3	7	10	44	7
Couldn't be improved	1	19	0	486	0	1	25	5	0
Dignity	0	4	0	0	9	0	0	5	0
Environment/ facilities	2	35	8	4	1	107	5	16	1
Miscellaneous	0	4	0	5	0	0	1874	0	0
Staff	8	106	16	7	5	15	4	617	3
Transition/coordination	4	10	3	0	4	1	4	7	14

The third and final way of assessing model performance is *accuracy per class*. Accuracy per class is simply, for each tag, the number of times that this tag has been correctly predicted, divided by the number of records under this tag. It is pretty much like the accuracy score, although it is for each tag rather than for all records in the dataset. For example, accuracy per class for “Transition/coordination” in the figure would be 14 / (14 + 3 + 1 + 7 + 7 +10) = 33%.

Text mining dashboard

The text mining dashboard presents the results of the text classification model (predictions, model performance metrics and plots etc.), but also reports further analyses on the feedback text in order to surface potentially important information, including how patients feel about the service and what they mostly talk about.

A detailed, high-level description of what type of analyses the dashboard does and how these are presented on the dashboard can be found here https://github.com/CDU-data-science-team/pxtextminingdashboard#readme.

Challenges with sentiment analysis

Sentiment analysis can help surface invaluable information that can guide managers in improving services. There are, however, a few drawbacks that need to be discussed here.

Sentiment analysis can sometimes fail to fully capture the true feelings expressed in a text. It is particularly hard to detect negation, colloquialisms, irony and sarcasm, among others. Moreover, patients often express a blend of both positive and negative sentiments, making it hard to summarize this into a sentiment “score”. A real example of the challenges of sentiment analysis is comment “I would like to thank you for working in these bad times.”, which a sentiment analysis algorithm may consider as negative because of the word “bad”. The sentiment outcome for this sentence pretty much depends on the sentiment analysis algorithm: some algorithms will simply split the text into words, get rid of words that have no sentiment value (e.g. stop words) and then count the number of positive or negative words. This type of algorithm would fail in the example above. Other algorithms are more sophisticated- they are built on Machine Learning models that are trained on massive amounts of comments that are available on the web, e.g. Twitter. Essentially, these models are trained to “understand” language. However, they may still fail to capture the true sentiment, because the context in which they are trained (e.g. Twitter tweets) may be significantly different from text that is about a particular and more specialized area (e.g. patient feedback) where we would like to apply the model. In any case, the benefits of sentiment analysis outweigh the drawbacks, simply because it brings to surface so much potentially useful information that it enhances the discoverability of issues in healthcare provision.

Summary of progress

Broadly, the aims of the project were:

- Develop machine learning algorithms that can accurately classify patient feedback according to its theme and criticality
- Produce data visualisation and reporting tools in order to allow non technical users to read and explore the themes and content of their patient feedback
- Engage pilot and rollout sites so as to ensure that the final product meets their needs and the needs of other similar users across the system.

The degree to which each of these aims was met will now be considered. For a complete list of project aims see Appendix 1.

Success of machine learning algorithms

There is extensive detail in online resources about the dashboard and the text classification pipeline, in particular:

- Pipeline installation. https://github.com/CDU-data-science-team/pxtextmining#installation
- Pipeline description. https://github.com/CDU-data-science-team/pxtextmining#pipeline
- Pipeline function documentation. https://cdu-data-science-team.github.io/pxtextmining/index.html
- Dashboard structure. https://github.com/CDU-data-science-team/pxtextminingdashboard#dashboard-structure

Text classification pipeline

Briefly, the pipeline was fitted with a random search that randomly tries different tunings of (hyper)parameters. Several models that are known to perform well in text classification contexts were tried out, namely, Logistic regression, linear Support Vector Machines (SVM), Random Forest, Bernoulli Naive Bayes (NB), Multinomial/Complement NB, Ridge, Perceptron and Passive-Aggressive. See the Scikit-learn API (https://scikit-learn.org/stable/modules/classes.html), but also the pipeline function documentation (https://cdu-data-science-team.github.io/pxtextmining/pxtextmining.factories.html#module-pxtextmining.factories.factory_pipeline).

The pipeline was frequently fitted as new labelled data were becoming available. At the time of writing, the pipeline was fit with a 5-fold cross-validation on 67% on the data (training set) with 800 repetitions for theme and 1200 repetitions for criticality, and the model selection metric was set to be class balance accuracy (Mosely, 2013). This setting was used for both response variables (theme and criticality).

Results for theme

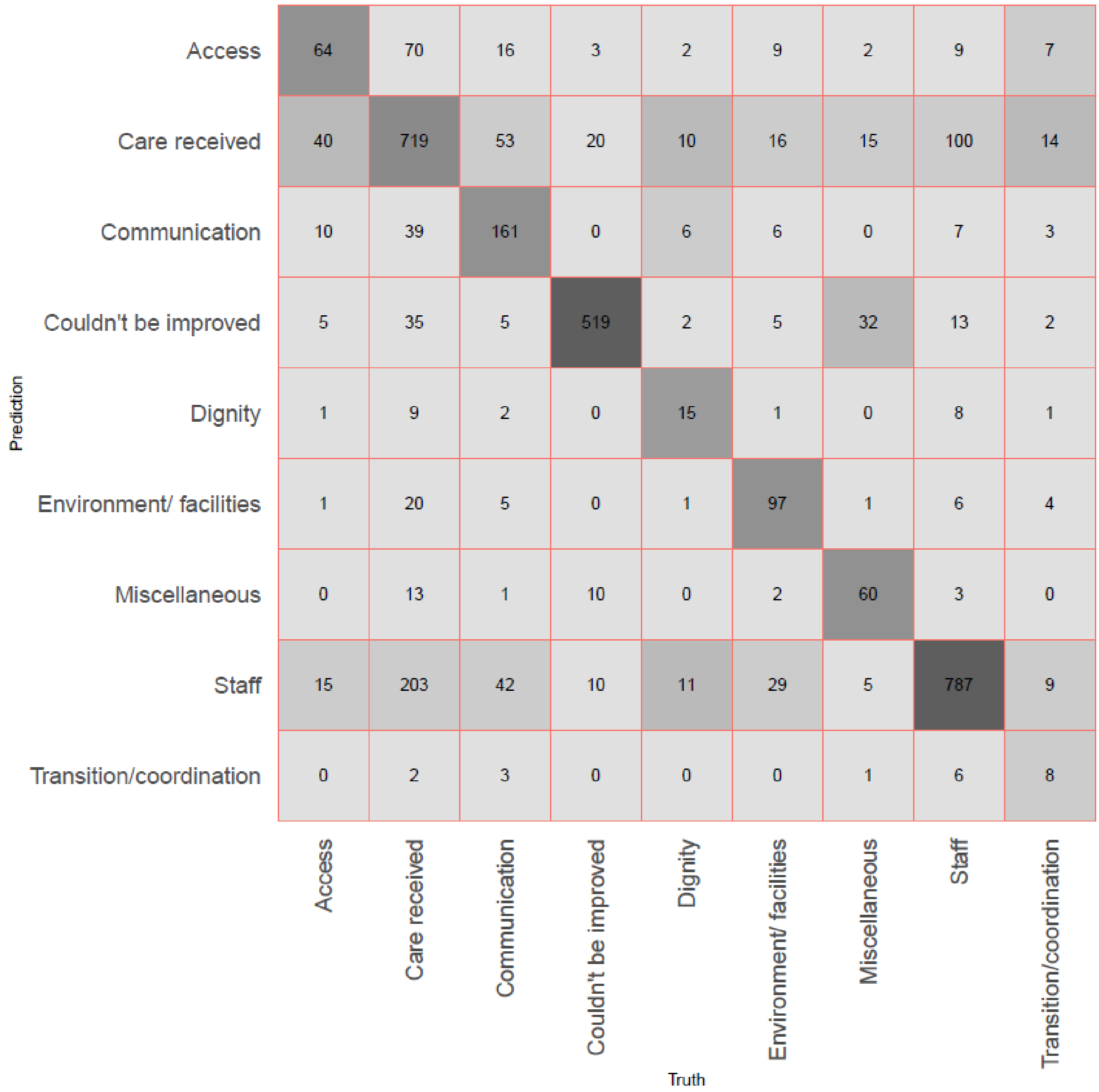
The winning model for the nine themes (see section “Number of tags”) was a linear SVM with 71% accuracy and 52% class balance accuracy on the test set.

Accuracy per class on the test set is reported as follows:

Theme	**Counts**	**Accuracy (%)**
Access	136	47
Care received	1110	65
Communication	288	56
Couldn't be improved	562	92

Dignity	47	32
Environment/ facilities	165	59
Miscellaneous	116	52
Staff	939	84
Transition/coordination	48	17

Confusion matrix on the test set (see “Assessing model performance”):



Prediction \ Truth	Access	Care received	Communication	Couldn't be improved	Dignity	Environment/ facilities	Miscellaneous	Staff	Transition/coordination
Access	64	70	16	3	2	9	2	9	7
Care received	40	719	53	20	10	16	15	100	14
Communication	10	39	161	0	6	6	0	7	3
Couldn't be improved	5	35	5	519	2	5	32	13	2
Dignity	1	9	2	0	15	1	0	8	1
Environment/ facilities	1	20	5	0	1	97	1	6	4
Miscellaneous	0	13	1	10	0	2	60	3	0
Staff	15	203	42	10	11	29	5	787	9
Transition/coordination	0	2	3	0	0	0	1	6	8

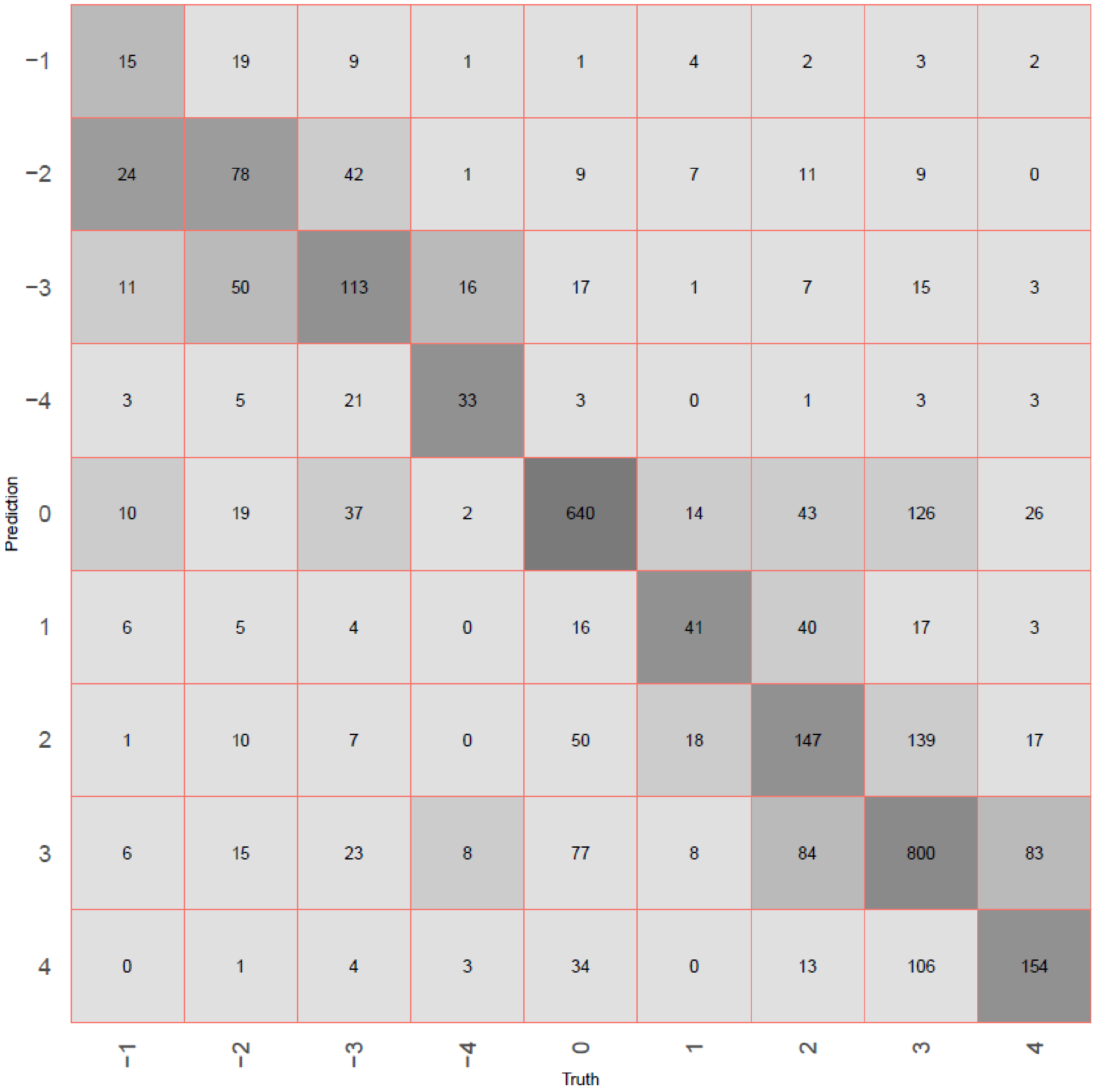
Results for criticality

The winning model for the nine criticality values (see sections “Number of tags” and “Lack of data”) was a logistic regression with 59% accuracy and 44% class balance accuracy on the test set.

Accuracy per class on the test set is reported as follows:

Criticality	Counts	Accuracy (%)
-1	76	20
-2	202	39
-3	260	43
-4	64	52
0	847	76
1	93	44
2	348	42
3	1218	66
4	291	53

Confusion matrix on the test set (see “Assessing model performance”):



	−1	−2	−3	−4	0	1	2	3	4
−1	15	19	9	1	1	4	2	3	2
−2	24	78	42	1	9	7	11	9	0
−3	11	50	113	16	17	1	7	15	3
−4	3	5	21	33	3	0	1	3	3
0	10	19	37	2	640	14	43	126	26
1	6	5	4	0	16	41	40	17	3
2	1	10	7	0	50	18	147	139	17
3	6	15	23	8	77	8	84	800	83
4	0	1	4	3	34	0	13	106	154

Text mining dashboard

Sentiment analysis and analysis of word frequencies that are presented on the dashboard (https://github.com/CDU-data-science-team/pxtextminingdashboard) are heavily based on R package tidytext (Silge & Robinson, 2017), for which a dedicated R package was developed, called experienceAnalysis (https://github.com/CDU-data-science-team/experienceAnalysis). Further sentiment analysis was performed with Python's TextBlob, for which there is an R wrapper function in package pxtextmineR (https://github.com/nhs-r-community/pxtextmineR).

The dashboard was built with shiny (https://shiny.rstudio.com/) using golem (https://thinkr-open.github.io/golem/index.html), the latter being a framework that promotes a modular, documented and testable, shareable and agnostic-to-deployment approach to development.

Evaluation

The degree to which the work, including the reporting and visualisation and reporting tools, engaged and met the needs of the pilot and rollout sites was considered by conducting semi-structured interviews with:

- Nottinghamshire University Hospitals NHS Trust (partner organisation)
- East Lancashire Hospitals NHS Trust (early adopter)
- Hereford and Worcestershire NHS Trust (early adopter)

The evaluation focused on the reporting dashboard *experiencesdashboard* https://github.com/CDU-data-science-team/experiencesdashboard as opposed to the more technically oriented *pxtextminingdashboard* discussed above which was used to generate insights about the performance of the models which were generated. The Trusts were encouraged to return to the programme objectives to reflect on their understanding of these and their perspectives on whether the programme achieved, and whether the software developed was of value to them and their organisation. The interviews focused predominantly on objective 1 and 3 (see Objectives in Appendix 1).

Perspectives on the programme purpose and ambitions

There were varying levels of prior understanding of machine learning and NLP in the Trusts involved. Those working in data science/analytics roles recognised potential in the ambition to create an algorithm to analyse and tag patient comments, those working in patient experience roles focussed more on the outputs and the resource associated with manual analysis. In some cases, interviewees stated that their organisations had no means to analyse comments and therefore comments were either ‘ignored’ or analysed manually.

One interviewee felt that the programme at one stage veered away from a focus on FFT data, towards other survey data. This was raised and while both FFT and survey data are identified within the programme objectives, the interviewee felt that their focus on FFT was understood and prioritised.

At numerous points, interviewees talked about the desire to ‘do more’ with comments. Those working in patient experience roles commented that this was useful data and would form part of how they work with services to use feedback to improve care. It was not clear whether the Trusts involved placed emphasis or value on patient comments specifically (beyond the interviewee’s interest). This was not mentioned and the absence of this within the interviews might suggest that the insight within patient comments is not widely recognised as valuable data, regardless of how effectively it can be analysed and reported.

All Trusts expressed a strong need for better data visualisation and reporting to support the use of patient experience feedback, aligning with objective 3. This came through as a priority to all interviewees, particularly those who worked closely with service teams and felt that currently they were unable to convince staff of the value of comments due, in part, to the difficulties with reporting comments succinctly and showing trend.

All Trusts felt there was real benefit from the project initiating within the NHS, led by people with an understanding of patient experience, of how patient comments are captured and of how patient feedback tends to be reported (and to whom).

All three of the Trusts stated that they appreciated that the programme was run from an NHS Trust, with the intention to release the software at no/low cost to NHS Trusts. Two of the Trusts spoke of NHS values and one interviewee made a comment relating to trust, stating that they felt able to commit to this programme and to contribute to its development because the origins and objectives felt legitimate.

Two of the Trusts viewed the programme as a valuable starting point, from which they (optimistically) hoped a useful solution would result following some further work. One of the Trusts felt that the solution was already creating benefit in releasing staff time (from manual analysis) to concentrate on working with services to use the comment data more effectively.

All interviewees stated that they would gladly remain involved in the development of the solution.

Experience of alternative commercial solutions

All Trusts were currently exploring or had explored commercial options for feedback analysis and reporting. All of the commercial options were viewed as expensive and in each case, the interviewees didn't feel that the commercial options provided the desired outputs. Two Trusts identified one of the issues as the disconnect between software development and the delivery of NHS care; they felt that in the case of commercial companies, specialised in software development and failed to understand what was needed.

One Trust stated that the software developed by this programme was 'significantly more sophisticated' from a data science perspective, and that the commercial solution their Trust had procured 'could not compete' (in terms of the accuracy and usefulness of outputs). This Trust found the analysis and reporting 'more crude in its analysis' and intended to undertake a 3-way comparison of the outputs of manual comment analysis, and comment analysis by this software and the software from the commercial supplier (this has not, at the time of writing, been completed).

The two Trusts currently exploring alternative commercial solutions expressed frustration at these solutions having been procured by staff/panels who they felt had not accurately judged the need of service staff. They felt that the solutions, while costly, had not delivered what was needed and were largely unnecessary and unhelpful. However, in both cases the interviewees felt that the financial commitment to these commercial systems had hindered their contribution to this project and reduced the resource they could commit to it.

The interviewees felt that the dashboards provided by the commercial alternatives had preferable user interfaces and identified this as an area of improvement within continued work.

The algorithm

Thematic tagging: All Trusts found the thematic tagging useful and accurate. One Trust stated that they were reassured about accuracy given that the system can learn/improve.

One Trust felt strongly that for the thematic tagging to be most useful and enable onward reporting, the themes/categories would need to be revised to better align with acute services (e.g. geography not relevant as based from a small number of large sites, but navigation of sites might be much more relevant).

In every case, the Trusts had read through the comments as categorised by the algorithm and stated that they could understand why the comments had been categorised as such.

Two Trusts stated that they wanted to see a more granular level of thematic tagging (e.g. subcategories). One Trust stated that this would ‘be enough to tip the balance’ from the commercial product to this software, enabling a case to be made that this software was preferable, and that the detail would secure better engagement from service teams. One interviewee is involved heavily in instigating service improvement work and felt that the additional granularity would pinpoint areas for improvement and give a sharper focus the work.

One Trust queried whether a comment which covered multiple themes would appear under each searchable element, or just one. This was evidently not clear in the explanation of the dashboard functions. Given a single tagging approach, the latter would apply, but this needs to be made clearer and documented if the situation should change due to a move to multiple tagging.

Sentiment categorisation: All Trusts found this useful and sufficiently accurate, though all Trusts stated that sentiment and criticality judgements were not as accurate as thematic tagging.

Two Trusts identified a specific issue with comments written with negative language but positive intent (e.g. *‘the service couldn’t do more’*); they had identified comments of this nature which had been incorrectly deciphered as critical.

Criticality categorisation: All Trusts found this helpful but currently found it insufficiently accurate to trust. There was recognition from two Trusts that judgement on criticality is very difficult and though this would be very valuable, they had not expected this programme to achieve this.

Two of the interviewees, those in data science roles, shared that they had experience of sentiment tagging tools and had not experienced anything as highly accurate yet.

All three Trusts identified that the most useful element of sentiment/criticality categorisation was the ability to quickly identify the most critical comments, which they consider to be the comments which require attention and might highlight serious issues of care quality. They considered discoverability of very critical comments highly valuable but felt that currently the judgements being made by the algorithm were not sufficiently reliable to depend on.

One Trust stated that criticality could be one of the most useful aspects of the software if the visualisations were more intuitive and showed quickly and simply how issue category frequency was changing over time.

The dashboard

All three Trusts initially stated that a user didn’t need to have a high level of technical ability to use the dashboard, stating that it was easy to navigate, intuitively designed, well set out and one Trust appreciated that searches and results loaded quickly when selecting fields/timeframes.

All Trusts interviewed had little difficulty uploading data to the software and stated that this was not an issue. One Trust had to reformat data to fit but didn't see this as a significant issue.

All the Trusts valued the options to search by timeframe, theme etc, and could see value in additional fields (e.g. specific services, type of feedback – if they ran other comment data through the software). One Trust were particularly complimentary of the way the demographic information was presented.

The Trusts did however identify various areas for improvement to enhance the user interface and the easy of navigation/understanding:

Graphs/visualisations:

- Need more explanation on the axis and in header text (particularly Sentiment Combinations). The interviewees had been able to make sense of most of the graphs but it had required considerable attention which is not something they feel service-level staff will commit to.
- One Trust talked about the visualisations needing to be 'instantly understandable'.
- The colours used in some visualisations didn't follow a linear gradient which would have made them easier to comprehend quickly.

Trend data:

- This information is particularly useful (all three Trusts mentioned this, and stated that this is the kind of information they are asked to report). One Trust felt that it was on good trend visualisation that the software would be sold to service-level and Board-level staff.

Summary and detail:

- One Trust stated that a preferable presentation would be to have a summary page, a headline chart, a single entry point on which patient experience teams could 'hook in' service-level staff. They appreciated the level of detail that could be surfaced (and wanted this in their own role) but felt that this was 'too much' for most people – that most people would engage with the software on the basis of 2-3 graphs, a summary or a single report.

Outputs:

- One Trust requested alternative output options (e.g. PDF, PowerBI)

Input and collaboration

One Trust spoke of the balance and pace of information and progress sharing. They were enthusiastic about understanding the data science underpinning the algorithm and the decisions being made relating to thematic categorisation but didn't feel they were able to engage satisfactorily with this. They expressed some disappointment with being told at times that detailed documentation/explanation was to follow, and at other times being given an opportunity to understand more but feeling overwhelmed with the explanation (which they stated were 'extraordinary, and significantly more impressive than [the commercial solution their Trust have procured]' but at times a little 'too much').

They didn’t feel they had ample opportunity to interrogate decisions/processes and as a result don’t feel that the solution fully accommodates their needs. That being said, the Trust remain very much committed to any continued work and were pleased to be involved. They were clear that the frustration came from wanting to contribute and understand, not from any lack of belief in the solution or the work in the background.

The two other Trusts stated that they felt sufficiently involved and updated on project progress and felt satisfied by their experience of working with Notts Healthcare (specifically the project lead).

All three Trusts were highly complementary of the project lead’s knowledge and expertise. This was particularly true of those in data science/analytics roles. All three Trusts spoke at length about their interactions with the project lead, having very clearly appreciated how responsive they were to them.

One Trust stated that they felt confused about the role of NHSE, that part of the value to them in engaging with the project was that it was centrally funded and supported and they have expected more input from NHSE so that they could understand the national direction, the ambitions for this project and comment analysis in general. They felt that this was a missed opportunity to triangulate expertise and perspectives.

One Trust stated that the project needed more structure and management. They felt that the project lead was holding everything together and would have appreciated more communication and project management. They were conscious of overwhelming the project lead.

Two Trusts stated that they would like to have more interaction with other Trusts using this software and they felt that they would glean useful insights from this. One Trust mentioned the value of an expert reference group/user panel.

One interviewee stated that they were glad they had involved some technical at their end (particularly to understand elements of IG, Cloud servers and software), and recommended this for any organisation wanting to use the software.

Value of the software/dashboard

Beyond the previously mentioned value of reducing human resource to manually analyse comments, two Trusts mentioned that the software would free them (particularly patient experience staff) to concentrate on outcomes and improvements as a result of feedback, rather than on the data management and feedback analysis.

As per aforementioned comments relating to commercial solutions, two Trusts were currently engaging with a commercial company on patient experience data and one Trust had experience of doing so previously. All three Trusts stated that their organisations would be in a position to pay for the software if a cost was associated with it, and two Trusts said that they would be happy to submit the case for funding immediately if that was required (with the second stating that they would want to do this on the condition of continued development, in which they could be involved).

Two Trusts stated that their organisations had made a substantial financial commitment to a commercial solution which they were 'wedded to' in a way they were not with this solution, despite it delivering more accurate outputs.

All three Trusts talked about added value coming from additional granularity of information, clearer trend data and more intuitive visualisations. It was on these elements that they felt the solution could distinguish itself from competitive alternatives.

All three Trusts mentioned bespoke/internal surveys (some Trustwide, some service-specific). All three Trusts felt that for maximum value, the software would incorporate and analyse comments from multiple feedback sources (including surveys and incident descriptions) and report on these in a coherent way.

One Trust felt that currently the dashboard wasn't adding value beyond the functionality of the commercial alternative not because it wasn't preferable, but because the Trust wasn't sufficiently invested in reading/ understanding/ using comments and therefore added sophistication in analysis/reporting on comments wasn't valued.

One Trust had a strong ambition to use the software to surface positive comments in the spirit of using these to share good practice and instigate quality projects from what the services were already doing well.

Continued development/requested features

All three Trusts stated that they would be willing to be involved in further development. The Trusts identified numerous developmental directions/features that would add value (some of which are mentioned in earlier sections of this report):

- Multiple theme tagging
- Wordclouds (mentioned both negatively, as very basic, and positively, as a way some people would want to engage with the data
- Additional data visualisation (e.g. SPC charts, funnel plots, advanced filtering options)
- Clearer, more engaging visualisation (e.g. colour gradients on graphs, simple trend graphs)
- More accessible and intuitive design to the dashboard (including narrative and explanation, tool tips etc)
- Ability to analyse patient comments from other sources
- Space to display how the services have acted upon the data - to keep the data and the resulting changes in one place
- Option to 'plug' outputs into other dashboards to build some curiosity for this information where staff are already engaged (e.g. PowerBI).
- Ability to click on an individual comment (or portion of a comment) and track the comment back to the service it relates to, or the other data this person had shared.

Summary

This project set out to produce free and open source tools for classifying patient feedback as well as software that would help users to explore, visualise, and report on their feedback after it had been classified by the model. As described in this report, these goals were achieved with ML metrics as well as evaluative methods indicating that the sites who were involved in the work were satisfied that the model is accurate and that the reporting tools were useful. Further work should focus on further enhancing the usability of the system according to the demands of the partner organisations, in particular adding multiple tagging, adding in other sources of feedback, and various improvements to the design and usability of the dashboard.

Chris Beeley, Amy Gaskin-Williams, Andreas Soteriades

Nottinghamshire Healthcare NHS Foundation Trust

March 2022

Appendix 1: Aims and objectives of the programme

(extracted from the Memorandum of Understanding, signed May 2020)

The aims and objectives of this programme are as follows:
This project aims to improve the use of FFT and patient experience survey free text in selected trusts, working towards generating from this piece of work a national "support or guidance toolkit" to help drive service improvements.

This will be achieved through creating text mining software, originating in the NHS, which analyses qualitative patient experience data for theme and sentiment, displays this on an online dashboard and is freely available to all NHS Trusts.

The aim is to create and validate (through piloting in trusts) the software with data from Notts Healthcare, two further Trusts and Care Opinion. The solution will then be deployed (with support) to an additional three Trusts within 12 months.

Objectives:

1. Improve the processing and analysis of FFT and patient experience survey free text data using text analytics (e.g. machine learning), through creating, developing and deploying text mining software.

2. Develop a process and software that is reproducible, sustainable and can be easily implemented in different NHS provider organisations and services, i.e. reusable across the system at low costs using open source components.

3. Establish data visualisation and/or reporting approaches that support the use of patient experience feedback for quality improvement.

4. Gain a better understanding of the variation in NHS trust needs across different trusts and service settings, thereby creating an easily transferable and adaptable solution.

Appendix 2: List of theme and criticality codes

Themes

Access
Care received
Communication
Couldn't be improved
Dignity
Environment/ facilities
Miscellaneous
Staff
Transition/coordination

Criticality codes

-4, -3, -2, -1 (strongly critical to mildly critical)

0 Neither critical nor positive

+1, +2, +3, +4 (mildly positive to strongly positive)