Identifying Patient Experience from Online Resources via Sentiment Analysis and Topic Modelling

Mohammed Bahja
Department of Computer Science
Brunel University London
London, UK
Mohammed.bahja@brunel.ac.uk

Mark Lycett
Department of Computer Science
Brunel University London
London, UK
Mark.lycett@brunel.ac.uk

Abstract—**Positive patient experience is crucial for retaining patient loyalty and in understanding and acting upon limitations of the treatment or service provided. Online platforms, such as websites and forums, are excellent sources for collecting more reliable feedback, as they provide anonymity and ease of use to the patients. Information from online sources can be vast and unstructured, thereby making patient feedback analysis challenging. Recent advancements in text mining and sentiment analysis approaches can enable automated and granular analysis of patient feedback. In this paper, we present our research, for which we applied and evaluated text mining and machine learning models to a patient feedback database obtained from the NHS Choices website to predict the patient sentiment in the database. There were two iterations to our research. First, we applied a linguistic approach using machine learning and dictionary scoring algorithms to predict patient sentiment from patient feedback and the predicted sentiment was validated against the ratings provided by the patients in the database. Second, a topic modelling approach was applied to identify "themes" within patient feedback so as to understand better the nature of the associated sentiment score, thereby providing a richer understanding of patient opinion.**

Keywords: component; text mining, opinion mining, sentiment analysis, patient experience, patient feedback, NHS Choices

1. INTRODUCTION

Information and Communication Technologies (ICTs) have increasingly enabled health service providers to collect patient experience data anonymously through sources, such as websites, blogs, forums and even internet search queries. This approach automates the data collection method making it potentially less expensive and more convenient to the patient as well as facilitating easier storage [1-3]. Whilst the potential size, distribution and heterogeneity of information collected online poses challenges in interpreting meaningful outcomes [4], progress in areas of data mining and machine learning approaches (text mining in particular) offers promise in identifying patterns from a large pool of data and predicting or identifying salient information from the dataset [5].

The National Health Service (NHS) in the United Kingdom (UK) has an online portal called NHS Choices, where patients provide both ratings and reviews for a particular NHS clinic. The NHS Choices ratings system provides an overview of patient experience for a specific set of parameters, including 'cleanliness', 'dealt with dignity' and others. Through this approach, understanding patient feedback is partially constrained by the set of parameters in that it provides little information about other aspects of patient experience that may be of value. In this report on our research study, we use sentiment analysis and topic modelling approaches to review the textual patient feedback, thereby providing enhance understanding of patient experience.

The paper is structured as follows. First, a brief introduction to patient experience, sentiment analysis and topic modelling aspects is provided. Second, we describe the data pre-processing methods adopted in our research. Third, a description of the sentiment analysis methods applied to the patient feedback database is provided, alongside the results obtained from the application of these methods. Finally, we describe the topic modelling approach applied and discuss the results obtained.

2. PATIENT EXPERIENCE AND ANALYSIS

Patient feedback obtained from online sources is typically vast and unstructured. It is important, therefore, that such data be efficiently analysed to extract the maximum value. Various information analysis methods are available, which can be implemented on web/online-based patient feedback data so as to understand patient experience. Two approaches, sentiment analysis and topic modelling, are explored in our research in relation to analysing the patient feedback, which are briefly described next.

2.1 Sentiment Analysis

Sentiment analysis is a relatively recent information analysis approach, that is used to determine the attitude, emotion and opinion of the writer via machine-based analysis of textual data. Researchers typically use machine learning and statistical methods to identify and characterise the emotional content of a text [10]. Once the data are analysed, at the crudest level, they can be 'binned' into positive or negative (satisfied or dissatisfied) categories. The benefits of using this technique is that it automates the process, gives quick results when compared to manual approaches and is (arguably) free of human bias. Moreover, since this technique is used across a wide range of topics, it is a well-established procedure, meaning that trouble-shooting guidelines for errors and common pitfalls are widely available [11]. The challenges in using this approach

BDCAT'16, December 06-09 2016, Shanghai, China
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4617-7/16/12…$15.00
DOI: http://dx.doi.org/10.1145/3006299.3006335

generally relate to the complexity of the text, since humans can express opinions in different ways, e.g. sarcastic or ironic content can be misinterpreted [12]. Nonetheless, sentiment analysis is powerful and is being widely adopted for various applications.

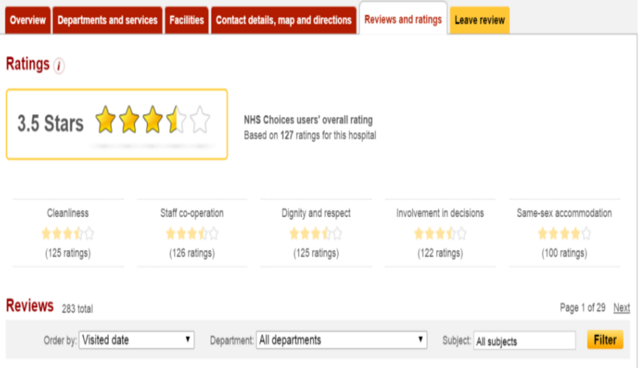
2.2 Topic Modelling

Topic modelling is a technique that identifies abstract ‘topics’ from a document that are useful for text categorisation and opinion mining [12]. The approach is based on the idea that a document is a mixture of topics, where each topic is a probability distribution in relation to words. For a given document, the distribution of words is identified and, from the clustering of words identified, a topic is derived from the document [13]. A topic modelling approach basically calculates the probability estimate of a word for a given topic, $P(w|t)$, and the probability of a topic for a given document, $P(t|d)$, for all the topics and documents analysed [14]. The approach’s advantages are that it enables the exploration of documents without a priori themes and uncovers the associations between different themes in the documents as well as how they evolve over time. Topic modelling can be used to annotate, summarise and organise large databases of electronic documents automatically, with minimal human intervention [15].

3. DATA PRE-PROCESSING

3.1 Patient Experience Data

In our first iteration of our research, a dataset of patient experience was collected from the NHS Choices website, relating to feedback on NHS hospital services. First, patients are asked to rate, on a scale of 1 to 5, how likely they are to recommend a particular hospital to family and friends. Second, they are asked to provide ratings on five parameters: (a) cleanliness; (b) staff co-operation; (c) dignity and respect; (d) involvement in decisions; and (e) same-sex accommodation. Third, the participants are given the option to provide a review on the hospital in their own words to a maximum of 3,000 characters. Figure 1 illustrates the review section from the NHS Choices website [16].



Ratings

3.5 Stars

NHS Choices users' overall rating
Based on 127 ratings for this hospital

Cleanliness	Staff co-operation	Dignity and respect	Involvement in decisions	Same-sex accommodation
(125 ratings)	(126 ratings)	(125 ratings)	(122 ratings)	(100 ratings)

Reviews 283 total

Page 1 of 29 Next

Order by: Visited date

Department: All departments

Subject: All subjects

Figure 1. A screenshot from the NHS choices website illustrating the various parameters of obtaining patient experience feedback (NHS Choices, 2016)

The dataset collected covers the ratings and feedback comments of patients who received healthcare from hospitals across the United Kingdom. The data collected cover the period between January 2010 and July 2015, comprising 76,151 comments with 56,818 labelled observations. Data analysis was performed using appropriate packages available in the R statistical software environment. R has been extensively used in previous research and has been shown to provide accurate classification results across different domains [17-19].

3.2 Data Processing

Within the dataset, the ratings given by the participants were used as the actual data against which the performance of the text mining model (i.e. in predicting the patient feedback) was tested. Those data were obtained from the ratings provided by the participants for the question, "How likely are you to recommend this hospital to friends and family if they needed similar care or treatment?”. The ratings for this question were considered in our tests, because this is the main question on the review webpage (other questions are optional) and the website uses the rating provided for this question for computing the overall recommendation rating of hospitals. Additionally, 78% of the participants have responded to rating of this question and thus, it has helped in assessing the overall patient experience.

Due to the skewed distribution of the numerical responses and limitations of the machine learning methods (and to reduce complexity in the modeling procedure), the continuous scale patient feedback ratings was discretised: A score of 1 or 2 were categorised as negative; a score of 4 and 5 were categorised as positive; and a score of 3 was discarded since there was no reliable way to categorise it as either positive or negative. This served as a binary sentiment label for which each text-mining model was trained and assessed.

After discretisation, a clean corpus was extracted from the reviews by applying functions in R to remove punctuation, reduce all letters to lower case and to deal with other minor formatting procedures, so as to create uniformity across the reviews. Consistency is crucial for model derivation, because small differences in words such as "Love" and "love" can reduce its predictive power. Once the corpus was created, the text was reformatted into a term document matrix, which is a simple, specialised structure that assigns a row to each document (review) and a column for each word. The cells corresponding to a document, word pair contain either a 0 or 1 indicating the non-presence or presence of the word in the document. Sparsity is a natural by-product of this specialised structure, which can negatively affect both computational and model performance. For example, rare words that occur in only one type of document, may be heavily weighted and bias the results. With certain approaches it is best practice to remove sparse words prior to analysis [21]. Here, we adopted a sparsity rule that a word needed to appear in at least five reviews to be considered for feedback classification.

4. RESULTS & DISCUSSION

4.1 Sentiment Analysis Models

In the first iteration of our research, our goal was only to classify the participant comment as binary category positive “1” or negative “0” (though later iterations of research will apply less crude sentiment categories and automatically mine classifications). To achieve binary classification, a linguistic approach of using a ‘bag of words’ was applied that identifies a group of words that may be associated with a sentiment. This was followed by the application of machine learning and dictionary scoring algorithms that predict the sentiment. The ratings provided by the patients in the actual data were then compared with the predicted ratings to assess the performance of the sentiment analysis models. The three different models applied were the: Strength of Association model (SoA); Support Vector Machine model (SVM); and the Naïve Bayes model (NB). Due to limited space, these models are not described in this paper, but overviews can be found in [5, 22, 23].

The three models were applied to the training, test and validation dataset, thereby being made fit to evaluate new comments (observations) for analysis. During the testing phase, the accuracy estimates of the model were observed and the model was tuned to improve the estimation accuracy of the prediction model. The performance validation of the SA models was carried out in two main stages. In the first stage, a single fold test dataset was used for validation. In the second stage, the dataset was divided over multiple folds as the test dataset and then used for cross validation of the model performance. Details of these two stages are provided below.

1) *Single fold Cross-Validation*

In this stage, the standard practice of performing sentiment analysis and prediction is used. The SA models read the dataset and learn to identify and classify the sentiment of the review, which is generally referred to as training the dataset. Further, the SA models are now trained for sentiment identification and classification. This implies that when a new review is presented to these trained SA models, they will be able to classify the sentiment. It is necessary for extensive testing of the model performance to be carried out and in our case, the test dataset was used. That is, the sentiment predictions of all the three models were performed on the test dataset. With a large dataset, along with the accuracy it is also essential to check the miscalculation delivered by the SA models. To this end, the prediction output of each model was collected in terms of true positives (TP), false negatives (FN), true negatives (TN) and false positives (FP). The accuracy and miscalculation estimates of the SA models for the test dataset are shown in the confusion matrix in Table 1.

Table 1. The sentiment prediction and miscalculation of the SA models for the test dataset. The overall prediction accuracy is also provided in the last column

SA model		Positive	Negative	Prediction accuracy (f-score)
SoA	**Positive**	5,967 **(TP)**	222 **(FN)**	0.67 (67%)
	Negative	4,346 **(FP)**	3,354 **(TN)**	
SVM	**Positive**	10,115 **(TP)**	1,945 **(FN)**	0.84 (84%)
	Negative	198 **(FP)**	1,631 **(TN)**	
NB	**Positive**	10,036 **(TP)**	2,715 **(FN)**	0.78 (78%)
	Negative	277 **(FP)**	861 **(TN)**	

In the above table, the true positives (TP), false negative (FN), true negatives (TN), and false positives (FP) predicted by each SA model is provided. For every instance where the sentiment predicted by the SA model for an individual review matches the ground truth sentiment of that review, then it is either a TP or TN depending on whether the sentiment is positive or negative. Thus, the TP, TN, FP, and FN for each SA model is obtained. In the last column of the table the prediction accuracy of each SA model is presented and it can be seen that the SVM model performs the best amongst the considered models, with a prediction accuracy of 84%. The NB approach has a prediction accuracy of 78%. The worst performance provided by the SoA approach, with an accuracy level of only 67%.

Further, the sensitivity and specificity of the models, were also calculated. Due to limited space, these aspects are not described in this paper, but overviews can be found in [25]. These two values were calculated using the following equations.

$$Sensitivity = \frac{number\ of\ TPs}{number\ of\ TPs + number\ of\ FNs}$$

$$Specificity = \frac{number\ of\ TNs}{number\ of\ TNs + number\ of\ FPs}$$

Using the above two equations, the sensitivity and specificity for all the three models were calculated and are shown in the table below.

Table 2. The sensitivity and specificity performance of the SA models

SA model	Sensitivity	Specificity
SoA	0.964	0.435
SVM	0.838	0.891
NB	0.78	0.756

In the above table, it can be seen that the sensitivity of the SoA model is the highest. In other words, despite the low prediction accuracy, it can be said that even at lower instances of identifying the positive sentiments, the SoA model was able to identify and classify the positive sentiments more precisely than the other models. The NB model has the lowest sensitivity of 0.78 and the SVM model has a sensitivity score of 0.838. Further, the SVM model has the highest specificity, i.e. the number of occasions of correct negative sentiments being identified by the SVM model is higher than for the other two SA models.

To visualize further the performance of the SA models in the study, segmented bar charts of the actual and predicted sentiment instances are provided in Figure 2. Prior to the visualisation, a summary of the ground truth number of positive and negative reviews along with the predicted values is shown in Table 3.

Table 3. The total number of positive and negative sentiment reviews identified by the three SA models along with the actual ground truth is provided

SA Model	Positive sentiment	Negative sentiment
Ground truth	10,313	3,576
SoA	6,189	7,700
SVM	12,060	1,829
NB	12,751	1,138

In the bar charts in Figure 2, it can be observed that the number of positive reviews identified by the SVM model is closer to that of the ground truth positive than the other models. However, it identified almost 40% fewer negative reviews when compared to the actual ground truth. The NB approach overestimated the number of positive reviews and also grossly underestimated the number of negative reviews present in the dataset. The SoA approach approximately identified almost 50% fewer positive reviews and estimated 50% more negative ones.

2) *Multi-fold Validation*

The performance of the model was further cross-validated by using a multi-fold cross validation approach for the study. Multi-fold validation is defined as the process of making multiple folds of the dataset as a test and training dataset. For instance, a five-fold cross validation study would involve dividing the dataset into ten different variations of training and test datasets, with the partition being randomly chosen [23].

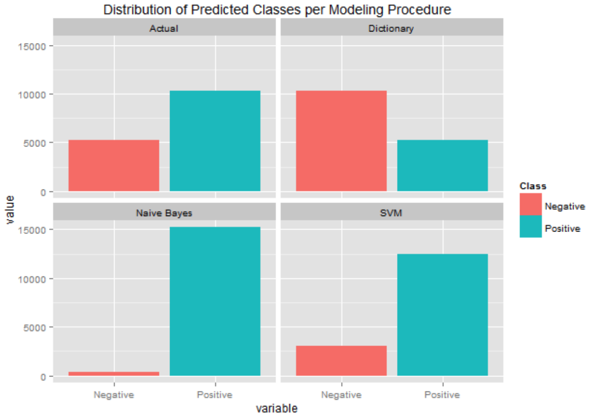


Figure 2. Plot illustrating performance of each SA model considered in our study against the "actual ground truth" data

In our study, the cross-validation was carried out with four folds, i.e. the dataset was partitioned into the training and test dataset four times and for each instance, the partition was performed randomly in *RStudio*. In both the single fold validation study and the four-fold one, the SVM model showed better performance than other two SA models considered in the tests. The SVM model provided high performance in terms of the prediction accuracy, i.e. the f-scores.

A summary of the f-scores, sensitivity and specificity for all the three SA models in both the single and four-fold validation studies is shown in Table 4. It can be observed that the SVM model performed better than other two models in terms of the f-scores in both single the fold and four-fold test. The NB approach shows similar prediction accuracies in both single and four-fold validation experiment. The SoA model has the lowest f-score in both instances of the validation study. In fact, the prediction accuracy of SoA diminishes in the four-fold validation experiment as compared to the single one.

Table 4. Summary of the f-scores, sensitivity and specificity scores of all the three models for both cross validation studies

SA Model		**F-scores**	**Sensitivity**	**Specificity**		
SVM	**Single-fold**	0.84	0.83	0.891		
	Four-fold	0.81	0.79	0.80		
NB	**Single-fold**	0.78	0.78	0.756		
	Four-fold	0.788	0.92	0.53		
SoA	**Single-fold**	0.67	0.96	0.435		
	Four-fold	0.603	0.62	0.78		

In terms of the sensitivity and specificity metrics, the performances were quite varied in both instances of the validation studies and it is difficult to point out a single SA model as being better than the others. In the single-fold study, the SOA metric provided the highest score of 0.96 and then followed by SVM and NB models, respectively. However, in the four-fold validation study, the NB model has a high sensitivity score of 0.92, SVM and SoA are 0.79 and 0.62, respectively. Similar varied performance in terms of specificity can also be seen for all the three models in both instances of the validation study. Whilst it is difficult to single out a particular model as the most suitable SA type for patient experience sentiment classification from these tests, the performance results indicate that the SVM provided the best performance when compared to other SA models. This result will need further testing and validation on different types of patient experience datasets and be extensively studied, for a stronger conclusion regarding the best patient sentiment classification model.

This study shows that the current SA models can be used to get an overview of the patient sentiments from a given dataset. It is desirable get a higher prediction accuracy than elicited in this study and by using a larger and more diverse dataset for training, the SA models can further improve their prediction accuracy. The prediction accuracy of almost 85% obtained by the SVM model is beneficial for the hospitals and clinics to understand that the identified sentiment has a 0.85 probability of being right. Further, the binary classification of the sentiment is the first stage or can be said to be a surface level sentiment analysis of the patient experience. The eventual sentiment analysis should be more fine-grained, being capable of identifying different aspects or features from the patient feedback database. It is this we now turn to address in the next section experiment.

4.2 Topic Modelling Approach

Identifying the sentiment in the patient feedback provides limited information. For instance, it would be helpful to know what topics are discussed in the feedback. We applied topic modelling methods to identify frequently occurring topics in the patient feedback – in essence to understand better the various aspects or nuances of the service and to be able to attribute the positive/negative sentiment to specific predicted topics.

Several topic modelling approaches exist, of which the Latent Dirichlet Allocation (LDA) is the most popular. The LDA approach assumes that each topic is a distribution of words and that each document has a certain distribution of topics. This assumption is further extended such that each word in the document belongs to or can be attributed to one of the topics contained within. The LDA approach identifies a theme or topic based on the probability distribution of a given set of words, which may belong to a certain theme. Further, it also uses the Dirichlet distribution to define the distribution of topics in a given document. A detailed description of LDA method can be found in [15, 26].

First, we applied LDA to the NHS choices patient feedback database, identifying 30 topics as an outcome. We 'tuned' the LDA approach to identify the top 25 words most likely to belong to a topic. For each topic, those 25 words were reviewed by hospital staff (i.e. domain expert), for verification and to provide a meaningful name for the topic.

When applying the approach, both unigram and bigram based modelling were deployed. The unigram modelling identifies a bag of words that probabilistically belongs to a topic by analysing the occurrence of individual words – e.g. such as 'birth', 'baby', 'midwives', 'pregnancy' etc – as belonging to a particular theme from the database. In this example case, the topic was named as Maternity Department. Conversely, the bigram modelling method categories a bag of words by looking for word 'pairs' that may belong to a particular topic – e.g. 'car park', 'parking space', 'blue badge', 'disabled-parking' etc – as belonging to a specific category. In this case, the topic was named Parking Infrastructure.

Once the topics were identified, the outcomes of each iteration were aligned. In doing this, for each topic, all the patient comments in which that topic was mentioned are identified along with the sentiment rating predicted by the SVM based sentiment analysis method described previously. The mean of the predicted sentiment score of all those comments is calculated to obtain the mean sentiment score of each topic. For instance, for the Maternity Department topic, the predicted sentiment scores of the comments in which the patients provided feedback about the maternity department service of a particular hospital was averaged to assess the feedback on the department.

Figure 3 presents a static screenshot of an interactive visualisation of the topic modelling results as applied to the dataset [27-28]. The right hand side of the figure shows the 30 most salient terms found in the patient feedback database: The left hand side of the figure are the 30 topics identified from the database and the area of each circle corresponds to the prevalence of the topic in the database. The centres of the circles are determined by computing the distance between topics and projected onto a two dimensional plane, as shown in the figure. This type of visualisation helps in understanding the extent with which the topics are discussed. The more a topic is discussed across the database, then the greater the area of circle for that topic. The distance between the circles is an indicator of how likely two topics are to be mentioned in a particular comment, which can facilitate analysis of the association between different topics.

In Figure 4, an illustration of the link between a topic and the terms belonging to it is shown. When a particular topic is selected, the most frequently occurring terms for that topic are visualised on the right side along with a comparison of its occurrence frequency in the entire database. In the figure, topic 9 is chosen, which corresponds to the "Maternity Department" topic and on the right side the frequently occurring terms in the database for that particular topic are shown. It can be observed that terms, such as baby, midwives, maternity almost exclusively correspond to the "Maternity Department" and these are indicated by the red bars. On the other hand, whilst terms, such as "hospital, staff, experience" may occur in the "Maternity Department", they also have a very high occurrence across the database as indicated by the blue bars.

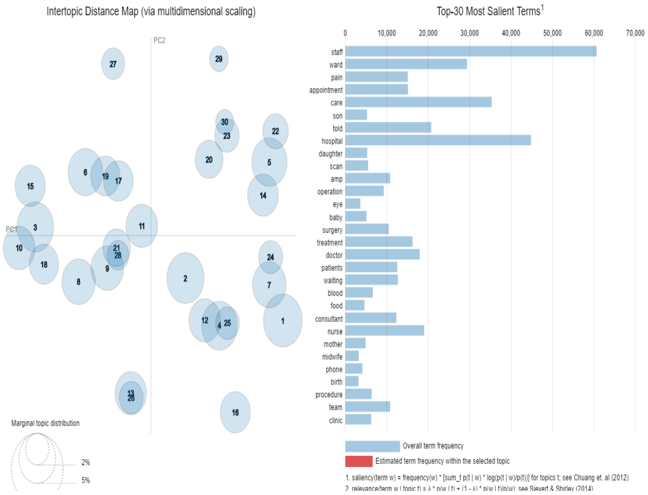

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
27
29
30
23
22
20
5
6
19
17
15
14
3
11
10
21
28
18
9
8
2
24
7
12
4
25
1
13
26
16
Marginal topic distribution
2%
5%
Top-30 Most Salient Terms1
0
10,000
20,000
30,000
40,000
50,000
60,000
70,000
staff
ward
pain
appointment
care
son
told
hospital
daughter
scan
amp
operation
eye
baby
surgery
treatment
doctor
patients
waiting
blood
food
consultant
nurse
mother
midwife
phone
birth
procedure
team
clinic
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. Visualisation of the 30 topics identified from the database on the left and the most frequently occurring terms in the database on the right

The next logical step after identifying the topics is to calculate the mean sentiment for each topic`s distribution over the database. The mean sentiment score gives an overall indication of the patient experience for the given topic across the dataset. Table 5 lists the 30 topics along with the mean sentiment score. In the table, it can be seen that certain services, such as Maternity Department and Ophthalmology achieved good sentiment scores, whereas the dental departments, gynaecology visits and delivery attained considerably lower ones.

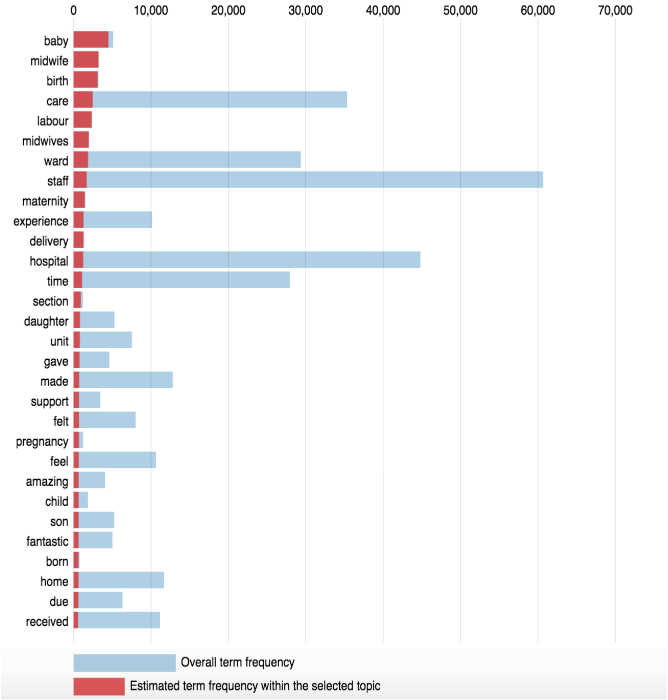


Figure 4. Visualisation of the topics, with topic 9 selected on the left

TABLE 5 MEAN SENTIMENT SCORES FOR THE 30 TOPICS IDENTIFIED BY THE UNIGRAM LDA MODEL

Topic	Mean Sentiment	Topic	Mean Sentiment
Maternity Department	4.715829	**Emergency Call/Delivery**	4.66
Emergency Room Service	2.129049	**Patient Reviews**	4.86
Dental Check-up	2.721543	**Ear Doctor Visits**	3.58
Knee Surgery	4.897586	**Family Doctor**	4.76
Gynaecology visits	2.83755	**General surgeon**	3.27
Hospital/Discharge Lounge	4.216019	**Cancer Treatment**	4.39
Waiting Room	2.021315	**Cardiovascular Treatment**	4.82

Ophthalmology	4.118987	**Inpatients Bathroom Complaints**	2.03
GI procedure	4.81401	**Operation Room**	4.87
Paediatric Visits	4.394958	**Elderly service department**	2.20
Patient Appointments and Follow-ups	1.978535	**Non-intensive Surgical Procedure**	3.08
Admission Ward	4.759339	**Hospital for Royal Families**	4.59
Clinic Service Experience	4.821004	**Delivery Room**	2.18
Patient's Good Experience	4.706137	**Parking Space Availability**	3.14
Telephone Service Department	2.061702	**Reviews on Hospital Service**	2.12

5. CONCLUSION

Understanding patient experience is increasingly important when evaluating the quality of healthcare provision. Here we have presented research-in-progress that aims to use automated approaches to understand the nuances of patient sentiment better. In this paper, we have presented two iterations of research. In the first, the performance of three approaches to sentiment analysis – Strength of Association, Support Vector Machine and Naïve Bayes – were assessed in the context of a dataset related to patient experience (drawn from the NHS choices website). The study involved performing a single-fold study, i.e. the dataset was divided into a training and test sections and then used for validation of the performance of the model. Next, the performance of the model was further cross-validated by using a multi-fold cross validation approach for the study. The SVM approach provided the most accurate prediction (85% accuracy) in assessing patient feedback as being positive or negative. For the second iteration of the research, topic modelling was employed to uncover frequently occurring topics that were 'hidden' within the a priori categorisation of the NHS dataset. This allowed for associating the predicted sentiment in the patient feedback to the topics discussed in the feedback. We make no major claims at this point aside from the observation that text-mining approaches can be effectively utilised to understand patient experience from online sources in a simple, inexpensive and efficient way. Our next iteration of research will build on the work here by exploring natural language processing and dependency parsing methods to analyse the reason behind the patient sentiment for a particular topic.

6. REFERENCES

[1] W. Jason, N. Victoria, M. Dianne and L. Sherri, '"Defining Patient Experience," *Patient Experience Journal*, vol. 1, no. 1, -04-30, pp. 7-19, April 2014.

[2] M. Beattie, W. Lauder, I. Atherton and D. Murphy, '"Instruments to measure patient experience of health care quality in hospitals: a systematic review protocol," *Systematic Reviews*, vol. 3, no. 1, pp. 4., 2014

[3] M.P. Manary, W. Boulding, R. Staelin and S.W. Glickman, '"The patient experience and health outcomes," *N.Engl.J.Med.*, vol. 368, no. 3, pp. 201-203.,2013

[4] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi and L. Donaldson, '"Use of sentiment analysis for capturing patient experience from free-text comments posted online," *J.Med.Internet Res.*, vol. 15, no. 11, Nov 1, pp. e239., 2013

[5] N. Collier, '"Uncovering text mining: A survey of current work on web-based epidemic intelligence," *Global public health*, vol. 7, no. 7, pp. 731-749., 2012

[6] F. Greaves, U.J. Pape, D. King, A. Darzi, A. Majeed, R.M. Wachter and C. Millett, '"Associations between Web-based patient ratings and objective measures of hospital quality," *Arch.Intern.Med.*, vol. 172, no. 5, pp. 435-436., 2012

[7] C. Lees, '"Measuring the patient experience," *Nurse.Res.*, vol. 19, no. 1, pp. 25-28., 2011

[8] A. Coulter, R. Fitzpatrick and J. Cornwell, '"The point of care measures of patients' experience in hospital: purpose, methods and uses," *The Kings Fund*, pp. 1-32., 2009

[9] S. Andrew, '"Evaluation and measurement of patient experience," *Patient Experience Journal*, vol. 1, no. 1, pp. 28-36., 2014

[10] B. Pang and L. Lee, '"Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1-135., 2008

[11] B. Liu, '"Sentiment analysis and subjectivity," *Handbook of natural language processing*, vol. 2, pp. 627-666., 2010

[12] R. Feldman, '"Techniques and applications for sentiment analysis," *Commun ACM*, vol. 56, no. 4, pp. 82-89., 2013

[13] M. Steyvers and T. Griffiths, '"Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424-440., 2007

[14] J. Uys, N. Du Preez and E. Uys, '"Leveraging unstructured information using topic modelling,", pp. 955-961., 2008

[15] D.M. Blei, '"Probabilistic topic models," *Commun ACM*, vol. 55, no. 4, pp. 77-84., 2012

[16] NHS Choices, '"Reviews and ratings - Hillingdon Hospital - NHS Choices,", Nhs.uk. N.p., 2016. Web. 15 Aug. 2016.

[17] Y. Zhao, '"R and data mining: Examples and case studies,", 2012.

[18] D. Meyer, K. Hornik and I. Feinerer, '"Text mining infrastructure in R," *Journal of Statistical Software*, vol. 25, no. 5, pp. 1-54., 2008

[19] I. Feinerer, '"Introduction to the tm Package Text Mining in R," 2013

[20] I. Feinerer, '"An introduction to text mining in R," *R News*, vol. 8, no. 2, pp. 19-22., 2013

[21] I. Feinerer, '"Introduction to the tm Package Text Mining in R," http://cran.r-project.org/web/packages/tm/ vignettes/tm.pdf., 2013

[22] P. Turney, '"Mining the web for synonyms: PMI-IR versus LSA on TOEFL,"., 2001

[23] E. Alpaydin, '"Introduction to machine learning,", 2014.

[24] K.L. Wuensch, '"What is a Likert Scale? and How Do You Pronounce'Likert?'," *East Carolina University*., 2005

[25] J.F. Hair, '"Multivariate data analysis,"., 2009

[26] D. Ramage, C.D. Manning and S. Dumais, '"Partially labeled topic models for interpretable text mining,", pp. 457-465., 2011

[27] J. Chuang,, C. D. Manning, and J. Heer., "Termite: Visualization techniques for assessing textual topic models.", *Advanced Visual Interfaces*, 2012.

[28] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics", pp. 63-70., 2014