BMJ Health & Care Informatics

Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review

Mustafa Khanbhai ,[1] Patrick Anyadi,[1] Joshua Symons,[2] Kelsey Flott,[1] Ara Darzi,[3] Erik Mayer[1]

To cite: Khanbhai M, Anyadi P, Symons J, *et al*. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform* 2021;**28**:e100262. doi:10.1136/bmjhci-2020-100262

Received 19 October 2020
Revised 03 January 2021
Accepted 12 January 2021

[1]Patient Safety Translational Research Centre, Imperial College of Science Technology and Medicine, London, UK
[2]Big Data and Analytical Unit, Imperial College of Science Technology and Medicine, London, UK
[3]Institute of Global Health Innovation, Imperial College of Science Technology and Medicine, London, UK

Correspondence to
Mustafa Khanbhai;
m.khanbhai@imperial.ac.uk

ABSTRACT

Objectives Unstructured free-text patient feedback contains rich information, and analysing these data manually would require a lot of personnel resources which are not available in most healthcare organisations. To undertake a systematic review of the literature on the use of natural language processing (NLP) and machine learning (ML) to process and analyse free-text patient experience data.

Methods Databases were systematically searched to identify articles published between January 2000 and December 2019 examining NLP to analyse free-text patient feedback. Due to the heterogeneous nature of the studies, a narrative synthesis was deemed most appropriate. Data related to the study purpose, corpus, methodology, performance metrics and indicators of quality were recorded.

Results Nineteen articles were included. The majority (80%) of studies applied language analysis techniques on patient feedback from social media sites (unsolicited) followed by structured surveys (solicited). Supervised learning was frequently used (n=9), followed by unsupervised (n=6) and semisupervised (n=3). Comments extracted from social media were analysed using an unsupervised approach, and free-text comments held within structured surveys were analysed using a supervised approach. Reported performance metrics included the precision, recall and F-measure, with support vector machine and Naïve Bayes being the best performing ML classifiers.

Conclusion NLP and ML have emerged as an important tool for processing unstructured free text. Both supervised and unsupervised approaches have their role depending on the data source. With the advancement of data analysis tools, these techniques may be useful to healthcare organisations to generate insight from the volumes of unstructured free-text data.

Summary

What is already known?

- The ability to analyse and interpret free-text patient experience feedback falls short due to the resource intensity required to manually extract crucial information.
- A semiautomated process to rapidly identify and categorise comments from free-text responses may overcome some of the barriers encountered, and this has proven successful in other industries.

What does this paper add?

- Natural language processing and machine learning (ML) have emerged as an important tool for processing unstructured free text from patient experience feedback.
- Comments extracted from social media were commonly analysed using an unsupervised approach, and free-text comments held within structured surveys were analysed using a supervised approach.
- Healthcare organisations can use the various ML approaches depending on the source of patient experience free-text data, that is, solicited or unsolicited (social media), to gain near real-time insight into patient experience.

BACKGROUND

Over the last decade, there has been a renewed effort focusing on patient experiences, demonstrating the importance of integrating patients' perceptions and needs into care delivery.[1 2] As healthcare providers continue to become patient-centric, it is essential that stakeholders are able to measure, report and improve experience of patients under their care. Policy discourse has progressed from being curious about patients' feedback, to actually collecting and using the output to drive quality improvement (QI).

In the English National Health Service (NHS), USA and many European health systems patient experience data are abundant and publicly available.[3 4] NHS England commissions the Friends and Family Test (FFT), a continuous improvement tool allowing patients and people who use NHS services to feedback on their experience.[5] It asks users to rate services, or experiences, on a numerical scale such as the Likert scale. In addition to quantitative metrics, experience

surveys such as the FFT also include qualitative data in the form of patient narratives. Evidence suggests that when staff are presented with both patient narratives and quantitative data, they tend to pay more attention to the narratives.[6] Patient narratives can even complement quantitative data by providing information on experiences not covered by quantitative data,[7 8] and give more detail that may help contextualise responses to structured questions. These free-text comments can be especially valuable if they are reported and analysed with the same scientific rigour already accorded to closed questions.[9 10] However, this process is limited by human resource and the lack of a systematic way to extract the useful insights from patient free-text comments to facilitate QI.[11 12]

Natural language processing (NLP) and machine learning (ML)

A potential solution to mitigate the resource constraints of qualitative analysis is NLP. NLP is currently the most widely used 'big data' analytical technique in healthcare,[13] and is defined as 'any computer-based algorithm that handles, augments and transforms natural language so that it can be represented for computation.'[14] NLP is used to extract information (ie, convert unstructured text into a structured form), perform syntactic processing (eg, tokenisation), capture meaning (ie, ascribe a concept to a word or group of words) and identify relationships (ie, ascribe relationships between concepts) from natural language free text through the use of defined language rules and relevant domain knowledge.[14–16] With regards to text analytics, the term ML refers to the application of a combination of statistical techniques in the form of algorithms that are able to complete diverse computation tasks,[17] including detect patterns including sentiment, entities, parts of speech and other phenomena within a text.[18]

Text analysis

Topic or text analysis is a method used to analyse large quantities of unstructured data, and the output reveals the main topics of each text.[19 20] ML enables topic analysis through automation using various algorithms, which largely falls under two main approaches, supervised and unsupervised.[21] The difference between these two main classes is the existence of labels in the training data subset.[22] Supervised ML involves predetermined output attribute besides the use of input attributes.[23] The algorithms attempt to predict and classify the predetermined attribute, and their accuracies and misclassification alongside other performance measures are dependent on the counts of the predetermined attribute correctly predicted or classified or otherwise.[22] In healthcare, Doing-Harris *et al*[24] identified the most common topics in free-text patient comments collected by healthcare services by designing automatic topic classifiers using a supervised approach. Conversely, unsupervised learning involves pattern recognition without the involvement of a target attribute.[22] Unsupervised algorithms identify inherent groupings within the unlabelled data and subsequently assign label to each data value.[25] Topics within a text can be detected using topic analysis models, simply by counting words and grouping similar words. Besides discovering the most frequently discussed topics in a given narrative, a topic model can be used to generate new insights within the free text.[26] Other studies have scraped patient experience data within comments from social media to detect topics using an unsupervised approach.[27 28]

Sentiment analysis

Sentiment analysis, also known as opinion mining, helps determine the emotive context within free-text data.[29 30] Sentiment analysis looks at users' expressions and in turn associates emotions within the analysed comments.[31] In patient feedback, it uses patterns among words to classify a comment into a complaint, or praise. This automated process benefits healthcare organisations by providing quick results when compared with a manual approach and is mostly free of human bias, however, reliability depends on the method used.[27 32 33] Studies have measured the sentiment of comments on the main NHS (NHS choices) over a 2-year period.[27 34] They found a strong agreement between the quantitative online rating of healthcare providers and analysis of sentiment using their individual automated approach.

NLP and patient experience feedback

Patient experience is mostly in natural language and in narrative free text. Most healthcare organisations hold large datasets pertaining to patient experience. In the Englanish NHS almost 30 million pieces of feedback have been collected, and the total rises by over a million a month, which according to NHS England is the 'biggest source of patient opinion in the world'.[5] Analysing these data manually would require a lot of personnel resources which are not available in most healthcare organisations.[5 35] Patient narratives contain multiple sentiments and may be about more than one care aspect; therefore, it is a challenge to extract information from such comments.[36] The advent of NLP and ML makes it far more feasible to analyse these data and can provide useful insights and complement structured data from surveys and other quality indicators.[37 38]

Outside of a healthcare organisation, there is an abundance of patient feedback on social media platforms such as Facebook, Twitter, and in the UK, NHS Choices and Care Opinion and other patient networks. This type of feedback gives information on non-traditional metrics, highlighting what patients truly value in their experiences by offering nuances that is often lacking in structured surveys.[39] Sentiment analysis has been applied ad hoc to online sources, such as blogs and social media[7 27 33 34] demonstrating in principle the utility of sentiment analysis for patient experience. There appears to be an appetite to explore the possibilities offered by NLP and ML within healthcare organisations to turn patient experience data into insight that can drive care delivery.[40 41] However, healthcare services need to be cognizant of what

NLP methodology to use depending on the source of patient experience feedback.[5] To date, no systematic review related to the automated extraction of information from patient experience feedback using NLP has been published. In this paper, we sought to review the body of literature and report the state of the science on the use of NLP and ML to process and analyse information from patient experience free-text feedback.

The aim of this study is to systematically review the literature on the use of NLP and ML to process and analyse free-text patient experience data. The objectives were to describe: (1) purpose and data source; (2) information (patient experience theme) extraction and sentiment analysis; (3) NLP methodology and performance metrics and (4) assess the studies for indicators of quality.

METHODS

Search strategy

The following databases were searched from January 2000 and December 2019; MEDLINE, EMBASE, PsycINFO, The Cochrane Library (Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, Cochrane Methodology Register), Global Health, Health Management Information Consortium, CINAHL and Web of Science. Grey literature and Google Scholar were used to extract articles that were not retrieved in the databases searched. Owing to the diversity of terms used inferring patient experience, combinations of search terms were used. The search terms, derived from the Medical Subject Headings vocabulary (US National Library of Medicine) for the database queries that were used can be found below. A review of the protocol was not published.

"natural language processing" OR "NLP" OR "text mining" OR "sentiment analysis" OR "opinion mining" OR "text classification" OR "document classification" OR "topic modelling" OR "machine learning" "supervised machine learning" OR "unsupervised machine learning" AND "feedback" OR "surveys and questionnaires" OR "data collection" OR "health care surveys" OR "assessment" OR "evaluation" AND "patient centred care" OR "patient satisfaction" OR "patient experience".

Inclusion criteria

To be eligible for inclusion in the review, the primary requirement was that the article needed to focus on the description, evaluation or use of NLP algorithm or pipeline to process or analyse patient experience data. The review included randomised controlled trials, non-randomised controlled trials, case–control studies, prospective and retrospective cohort studies and qualitative studies. Queries were limited to English language but not date constraints. We excluded studies that gathered patient-reported outcome measurements, symptom monitoring, symptom information, quality of life measures and ecological momentary assessment without patient experience data. Conference abstracts were excluded, as there was limited detail in the methodology to score against quality indicators.

Study selection

The research adhered to the guideline presented in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 checklist.[42] The initial search returned 1007 papers; after removing duplicates 241 papers were retained. The titles and abstract were screened by two reviewers (MK and PA) independently, and discrepancies were resolved by a third reviewer (EM). Thirty-one articles were identified as potentially eligible for inclusion. Full-text articles were retrieved and assessed for inclusion by the same reviewers, of which 19 were retained for final inclusion. The main reason for exclusion was the articles reported other patient-reported feedback and not patient experience. Figure 1 illustrates the PRISMA flowchart representing the study selection process and reasons for exclusion.

Data collection process

We developed a data collection tool with the following data fields: department of corresponding authors, country of study, study purpose, data source, solicited feedback, time period, information extraction method, data processing, ML classifiers, text analysis approach, software, performance, key findings and limitations. Two reviewers (MK and PA) independently completed the data collection, and met to compare the results, and discrepancies were resolved by a third reviewer (EM).

Data synthesis

Due to the heterogeneous nature of the studies, a narrative synthesis was deemed most appropriate. A formal quality assessment was not conducted, as relevant reporting standards have not been established for NLP articles. Instead, we report indicators of quality guided by elements reported in previous NLP-focused systematic reviews.[43–46] We included information related to the study purpose, corpus (eg, data source and number of comments), NLP (eg, methodology and software used and performance metrics). Two reviewers (MK and PA) independently evaluated indicators of quality in each study, disagreements in evaluation were resolved by discussion with a third reviewer (EM). Inter-rater agreement Cohen's kappa was calculated. In the reviewed studies, we assessed the NLP methodology and the rationale for its use. The key NLP approaches were summarised based on text analysis incorporating either text classification or topic modelling depending on the corpus available and evaluation was done as to whether sentiment analysis was performed using existing or bespoke software.

Performance metrics

To understand how well an automated ML algorithm performs, there are a number of statistical values that help determine its performance with the given data.[18] Algorithm performance is measured as recall (proportion of all true positive observations that are correct,

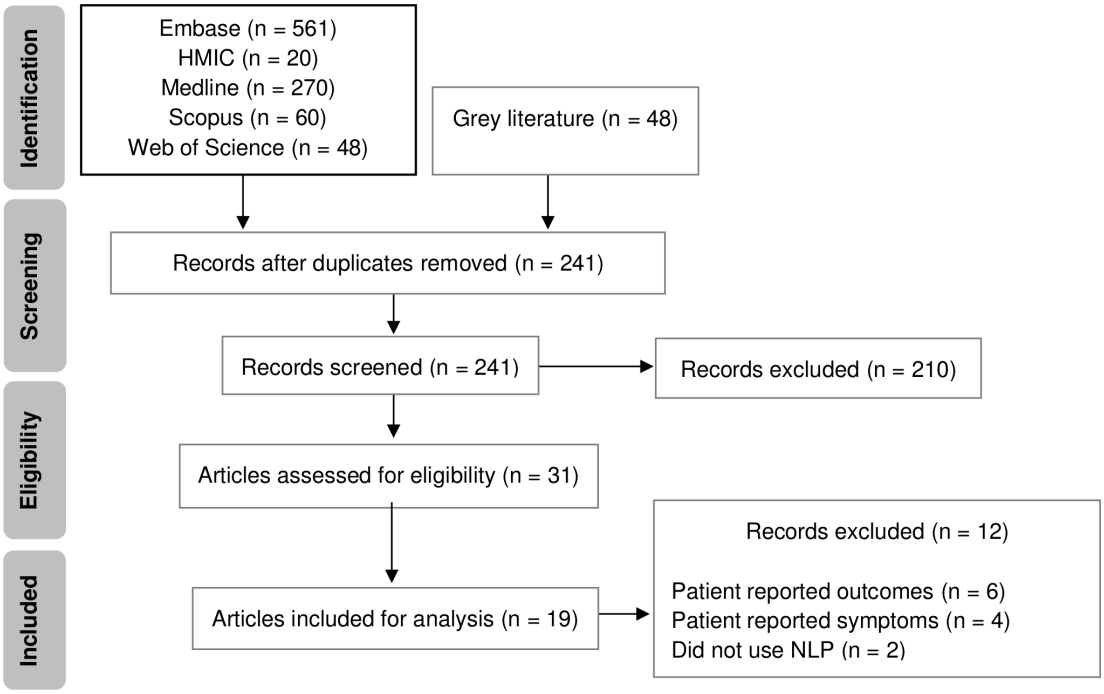

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 flowchart. NLP, natural language processing.

that is, true positives/(true positives+false negatives)), precision (ratio of correctly predicted positive observations to the total predicted positive observations) and by the F-score which describes overall performance, representing the harmonic mean of precision and recall.[43] K-fold cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. This ensures that the results are not by chance, and therefore ensures the validity of the algorithms performance. We look all the recorded performance metrics in each of the included studies in order to gain a better understanding of how the data and ML approach can influence the performance.

RESULTS

Study characteristics

Year of publication ranged from 2012 to 2020 with almost 80% (15/19) of articles published in the last 5 years. The study purpose of the 19 articles was similar, in that they applied language analysis techniques on free-text patient experience feedback to extract information, which included themes or topics and sentiment. The feedback was either solicited[24 47–50] or unsolicited.[6 26–28 32 34 51–58] Six studies were from the UK,[26–28 48 49 55] two from Spain,[58] of which one included Dutch reviews[54] and the rest were conducted in the USA,[6 24 32 34 47 50 52 53 56 57] of which one[51] looked at Chinese language reviews translated in English. The authors of all except one study[47] were from a healthcare informatics department.

Data source

The majority (15/19) of the feedback used for language analysis was extracted from social media sites, such as Twitter,[28 52] Facebook[6] and healthcare specific forums, for example, NHS Choices,[26 27 55] Yelp,[56 57] RateMDs,[32 34 53] Haodf,[51] Masquemedicos,[54 58] Zorgkaart Nederland.[54] RateMDs and Yelp are US platforms that provide information, reviews and ratings on everything from cleanliness of hospital and care centre facilities, to clinician knowledge, as well as giving patients the ability to share personal experiences of care. NHS Choices is a UK-based platform that allows patients, carers and friends to comment on their experience of care received in any NHS institution. Haodf, Masquemedicos and Zorgkaart Nederland are platforms that incorporate patient experiences in Chinese, Spanish and Dutch, respectively. Five studies used the accompanying free text from structured patient feedback surveys; Press Ganey,[24 50] vendor supplied (HCAHPS and comments),[47] bespoke cancer experience survey with free-text comments,[48] Cancer Patient Experience Survey.[49] The initial dataset in terms of number of reviews captured to perform language analysis varied significantly from 734 reviews[58] to 773 279 reviews.[51] Where provided, the number of words, characters or sentences within the reviews varied. Table 1 gives an overview of the length of comments provided as either range, mean or median.

Software

The most common coding environment, sometimes used in combination, was Python (n=5)[24 49 50 52 53] followed by R (n=3),[26 48 55] Waikato Environment for Knowledge

Table 1 The length of comments provided in five of the 19 studies, arranged in descending order according to the total number of comments

Author	Data source	No. of comments	Length of comments
Hao *et al*[51]	Haodf	773 279	Mean 95.75 characters
Rastegar-Mojarad *et al*[57]	Yelp	79 173	Median 635 characters
Wallace *et al*[32]	RateMDs	58 100	Median 41 words
Wagland *et al*[48]	Cancer survey	5636	1–225 words
Plaza-del-Arco *et al*[58]	Masquemedicos	734	Mean 44 words

Analysis (n=2),[27 34] Machine Learning for Language Toolkit (n=2),[53 56] RapidMiner (n=2),[6 58] and C++ (n=1).[54]

Language analysis approach

Studies used a variety of approaches to develop their language analysis methodology. The two most common approaches were supervised (n=9)[6 27 28 34 47 48 50 52 54] and unsupervised learning (n=6),[24 26 51 53 55 56] followed by a combination, that is, (semisupervised) (n=3),[32 57 58] rule-based (n=1)[49] and dictionary look-up (n=1)[54] (figure 2). Sentiment analysis with a combination of text analysis was performed in 10 studies,[24 26 28 32 47–49 52 53 57] sentiment analysis alone was performed in four[6 28 50 54] and text analysis alone in four studies.[51 55 56 58] We describe the details of the two approaches, sentiment analysis and text analysis, which incorporated text classification and topic modelling, categorised as supervised and unsupervised learning, respectively.

Supervised learning

Manual classification into topics or sentiment was performed in those studies that used a supervised approach. The most common approach was manual classification of a subset of comments as the training set. The percentage of total number of comments used for manual classification varied in each study, as did the number of raters. Sentiment was generally expressed as positive, negative and neutral. Five studies did not perform manual classification and employed existing software to perform the sentiment analysis, that is, TheySayLtd,[28] TextBlob,[52] SentiWordNet,[57] DICTION,[53] Keras.[50] We split the supervised approach based on sentiment analysis (table 2A) and text classification (table 2B), where we document the percentage of total comments manually classified into categories for sentiment and topics for text classification, the number of raters including the inter-rater agreement and the classifier(s) used for ML. In addition, where reported, we also highlight the configuration employed during the data processing steps. Support vector machine (SVM) was the most commonly used classifier (n=6) followed by Naïve Bayes (NB) (n=5).

Unsupervised learning

Topic modelling is an approach that automatically detects topics within a given comment. Seven studies[24 26 32 51 53 55 56] used this approach and majority of the studies (n=6)[24 26 51 53 55 56] used latent Dirichlet allocation (LDA). One study[32] used a variation, factorial LDA, however this was a semisupervised approach as it involved some manual coding. LDA is a generative model of text that assumes words in a document reflect a mixture of latent topics (each word is associated with a single topic). For the output to be understandable, the number of topics has to be chosen, and table 3 demonstrates the variation in topics determined while employing LDA.

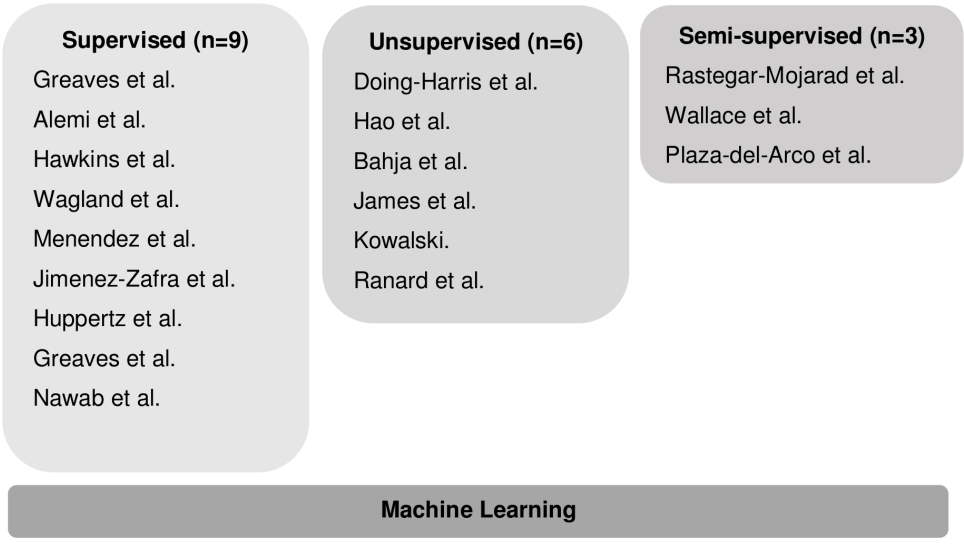

Figure 2 Most common approaches used to analyse free-text patient experience data identified in the systematic review.

Table 2A Studies that performed sentiment analysis using supervised approach, including the number of raters and associated inter-rater agreement expressed as Cohen's kappa (κ), classifiers and configuration applied where reported. Studies are reported in chronological order

Author	Data source	Comments classified	No. of raters	κ	Sentiment categories				Classifier							Configuration
					Positive	Negative	Mixed	Neutral	SVM	NB	DT	B	RF	GL	KN	
Alemi *et al*[34]	RateMDs	100%* (n=955)	NR	NR	✓	✓	✓	✓	✓	✓	✓	✓				Sparsity rule Information gain SVM: RBF kernel
Greaves *et al*[7]	NHS choices	17.56%† (1000/5695)	2	0.76	✓	✓	✓	✓	✓	✓	✓	✓				Prior polarity Information gain SVM: RBF kernel
Wagland *et al*[48]	Cancer experience	14.19% (800/5634)	3	0.64–0.87	✓	✓			✓			✓	✓	✓		NR
Bahja *et al*[26]	NHS choices	75% (56 818/76 151)	N/A	N/A	✓	✓			✓	✓						Sparsity rule Ratings in binary sentiment
Jimenez-Zafra *et al*[54]	COPOS and COPOD‡	100% (n=156975 COPOD and n=743 COPOS)	N/A	N/A	✓	✓			✓							Ratings in binary sentiment SVM: linear kernel
Huppertz *et al*[6]	Facebook	0.88% (508/57 986)	3	NR	✓	✓			✓	✓					✓	NR
Doing-Harris *et al*[24]	Press Ganey	0.58% (300/51 235)	3	0.73	✓	✓				✓						NR
Menendez *et al*[47]	Vendor supplied	100% (132/132)	NR	NR	✓	✓	✓	✓								NR

*Classified as praise (positive), complaint (negative), praise and complaint (mixed), neither (neutral).
†Only n-grams classified.
‡Also used dictionary lookup and cross domain method.
B, bagging; COPOD, corpus of patient opinions in Dutch; COPOS, corpus of patient opinions in Spanish; DT, decision trees; GL, generalised linear model; KN, k-nearest neighbour; NB, Naïve Bayes; NR, not reported; RBF, radial basis function; RF, random forest; SVM, support vector machine.

Table 2B Studies that performed text classification using supervised approach, including the number of rater and associated inter-rater agreement expressed as Cohen's kappa (κ), classifiers and configuration applied where reported. Studies are reported in chronological order

Author	Data source	Comments classified	No. of raters	κ	No. of themes	Classifier						Configuration
						SVM	NB	DT	B	RF	GL	KN	
Alemi *et al*[10 34]	RateMDs	100% (n=955)	NR	NR	9	✓	✓	✓	✓				Sparsity rule SVM: RBF kernel
Greaves *et al*[7]	NHS choices	*17.56% (1000/5695)	2	0.76	3	✓	✓	✓	✓				Prior polarity Information gain SVM: RBF kernel
Wagland *et al*[48]	Cancer experience	14.19% (800/5634)	3	0.64–0.87	11	✓			✓	✓	✓		NR
Doing-Harris *et al*[24]	Press Ganey	0.58% (300/51 235)	3	0.73	7		✓						NR
Hawkins *et al*[52]	Twitter	7511/11 602†	AMT	0.18–0.52	10	✓	✓						NR

*Only n-grams classified.
†Tweets classified as pertaining to patient experience only.
AMT, Amazon Mechanical Turk; B, bagging; DT, decision trees; GL, generalised linear model; KN, k-nearest neighbour; NB, Naïve Bayes; NR, not reported; RBF, radial basis function; RF, random forest; SVM, support vector machine.

Table 3 The number of topics arranged in descending determined in each study using latent Dirichlet allocation as a type of unsupervised learning approach

Author	Data source	No. of topics
Kowalksi	NHS choices	60
Ranard *et al*[56]	Yelp	50
Bahja *et al*[26]	NHS choices	30
Doing-Harris *et al*[24]	Press Ganey	30
Hao *et al*[51]	Haodf	10
James *et al*[53]	RateMDs	6

Performance

Seven studies did not report performance of the NLP algorithm or pipeline.[28 32 47 51 53 56 57] The remaining 12 studies reported one or more evaluation metrics such as accuracy, sensitivity, recall, specificity, precision, F-measure. The higher the F1 score the better, with 0 being the worst possible and one being the best. In the studies that employed a supervised approach, SVM and NB was the preferred classifier as it produced better results compared with other classifier demonstrated by the F1 score with sentiment analysis and text classification. Table 4 demonstrates the performance measure reported as F-measure or accuracy of the best performing classifiers for sentiment and text analysis using only supervised approach, and the k-fold cross-validation where reported in 12 studies, of which only five studies reported multiple fold validation.

Indicators of quality

Inter-rater agreement (Cohen's kappa) was calculated as 0.91 suggesting an almost perfect agreement. The individual evaluation with a description on each domain is detailed in table 5. Specifically, clarity of the study purpose statement, and presence of information related to the dataset, the number of comments analysed. information extraction and data processing, adequate description of NLP methodology and evaluation metrics. All studies had at least four of the seven quality indicators. Twelve studies addressed all seven indicators of quality,[6 24 26 27 34 48–50 52 54 55 58] and three studies addressed only four.[28 47 57]

DISCUSSION

In this systematic review, we identified 19 studies that evaluated various NLP and ML approaches to analyse free-text patient experience data. The majority of the studies dealt with documents written in English, perhaps because platforms for expressing emotions, opinions or comments related to health issues are mainly orientated towards Anglophones.[58] Three studies[51 54 58] were conducted using non-English free-text comments, however Hao *et al*[51] and Jimenez-Zafra *et al*[54] translated comments to English that were initially written in Chinese and Spanish, respectively. Accurate and automated analysis is challenging due to

Table 4 Performance metrics in the studies used supervised learning (sentiment analysis and text classification). SVM and NB were the preferred classifier as it produced better results demonstrated by the F1 score. Only five studies reported multiple fold validation

		Sentiment analysis		Text classification	
Author	**k-fold cross-validation**	**Classifier**	**Performance**	**Classifier**	**Performance**
Alemi *et al*[34]*†	Five repetitions of twofold cross-validation	SVM	Positive 0.89 Negative 0.64	SVM	Staff related 0.85 Doctor listens 0.34
		NB	Positive 0.94 Negative 0.68	NB	Staff related 0.80 Doctor listens 0.37
Doing-Harris *et al*[24]*	NR	NB	0.84	NB	Explanation 0.74 Friendliness 0.40
Greaves *et al*[27]	Single-fold cross-validation	NB SVM	0.89 0.84	NB SVM	Dignity and respect 0.85 Cleanliness 0.84 Dignity and respect 0.8 Cleanliness 0.84
Hawkins *et al*[52]	10-fold cross-validation	–	–	SVM	0.89‡
Jimenez-Zafra *et al*[54]	10-fold cross-validation	SVM	COPOD 0.86 COPOS 0.71	–	–
Huppertz *et al*[6]	NR	SVM	0.87‡	–	–
Wagland *et al*[48]	Single-fold cross-validation 10-fold cross-validation	SVM	0.80	–	–
		SVM	0.83	–	–
Bahja *et al*[26]	Single-fold cross-validation 4-fold cross-validation	SVM NB	0.84 0.78	–	–
		SVM NB	0.81 0.78	–	–

*Best and worst performing category, respectively.
†Classified as praise (positive), complaint (negative).
‡Reported as overall accuracy.
COPOD, corpus of patient opinions in Dutch; COPOS, corpus of patient opinions in Spanish; NB, Naïve Bayes; NR, not reported; SVM, support vector machine.

the subjectivity, complexity and creativity of the language used, and translating into other language may lose these subtleties. The type of patient feedback data used and choice of ML algorithm can affect the outcome of language analysis and classification. We show how studies used various ML approaches.

The two most common approaches were supervised and unsupervised learning for text and sentiment analysis. Briefly, text analysis identifies the topic mentioned within a given comment, whereas sentiment analysis identifies the emotion conveyed. Of the two approaches, the most common approach used was supervised learning, involving manual classification of a subset of data by themes[24 27 34 48 52] and sentiment.[6 24 26 27 34 48 52 54] Comprehensive reading of all comments within the dataset remains the 'gold standard' method for analysing free-text comments, and is currently the only way to ensure all relevant comments are coded and analysed.[48] This demonstrates that language analysis via an ML approach is only as good as the learning set that is used to inform it. The studies that used a supervised approach in this review demonstrated that there were at least two independent reviewers involved in manual coding, however, there was no consistency in the percentage of total comments coded, how the data was split into training and test set, and the k-fold cross-validation used. Within supervised learning, the most common classifier was SVM followed by NB. SVM and NB have been widely used for document classification, which consistently yield good classification performance.

NLP has problems processing noisy data, reducing overall accuracy.[18 59] Pre-processing of textual data is the first and an important step in processing of text that has been proven to improve performance of text classification models. The goal of pre-processing is to standardise the text.[59] We noted that pre-processing varied in the studies in this review. In addition to the standard pre-processing steps, that is, conversion to lowercase, stemming, stop word elimination, Alemi *et al*[34] used sparsity rule and information gain, Greaves *et al*[27] used information gain and prior polarity and Bahja *et al*[26] used sparsity rule alone. Plaza-del-Arco *et al*[58] used a combination of stopper and stemmer, and found that the accuracy was best (87.88%) with stemmer alone, however, F-measure was best (71.35%) when no stemmer or stopper was applied. However, despite these pre-processing steps, no

Table 5 Evaluation of studies and performance metrics

Author*	Defined purpose†	Data source described	Number of comments specified	Data processing described	Language analysis approach described	Evaluation metrics reported‡	Inclusion of comparative evaluation§
Alemi *et al*[34]	✓	✓	✓	✓	✓	✓	
Greaves *et al*[27]	✓	✓	✓	✓	✓	✓	✓
Greaves *et al*[28]	✓	✓	✓	✓			✓
Wallace *et al*[32]	✓	✓	✓	✓	✓		✓
Rastegar-Mojarad *et al*[57]	✓	✓	✓	✓			✓
Hawkins *et al*[52]	✓	✓	✓	✓	✓	✓	✓
Wagland *et al*[48]	✓	✓	✓	✓	✓	✓	
Hao *et al*[51]	✓	✓	✓	✓	✓		
James *et al*[53]	✓	✓	✓	✓	✓		
Bahja *et al*[26]	✓	✓	✓	✓	✓	✓	
Plaza-del-Arco *et al*[58]	✓	✓	✓	✓	✓	✓	
Doing-Harris *et al*[24]	✓	✓	✓	✓	✓	✓	
Huppertz *et al*[6]	✓	✓	✓	✓	✓	✓	
Kowalski	✓	✓	✓	✓	✓	✓	
Ranard *et al*[56]	✓	✓	✓	✓	✓		✓
Jimenez-Zafra *et al*[54]	✓	✓	✓	✓	✓	✓	
Menendez *et al*[47]	✓	✓	✓	✓			
Rivas *et al*[49]	✓	✓	✓	✓	✓	✓	
Nawab *et al*[50]	✓	✓	✓	✓	✓	✓	

✓ Indicates the presence of information in the article.
*Studies have been arranged in chronological order.
†Indicates reviewer judgement of clear statement of the study purpose.
‡Evaluation metrics include F-measure or accuracy.
§Comparison includes association with other survey data.

consensus could be found over a preferred supervised ML classification method to use for sentiment or text classification in the patient feedback domain.

The most interesting finding in this review was that the ML approach employed corresponded to the data source. The choice of approach is based on the performance metrics of the algorithm results, which depends on three factors.[21] First, identifying patterns is dependent on the quality of the data available. In text classification or sentiment analysis, the diversity of comments affects the accuracy of the machine prediction. More diversity decreases the ability of the ML algorithm to accurately classify the comment.[6] Second, each ML algorithm is governed by different sequential sets of rules for classifying semantic or syntactic relationships within the given text, and certain algorithms may suit some datasets better than others. Third, the larger the training sets used the higher the accuracy of the algorithms at identifying similar comments within the wider dataset, but trade-offs with time and human coding are necessary to ensure the method is resource-efficient.[21] We found that comments extracted from social media were commonly analysed using an unsupervised approach[26 32 51 53 55 56] and free-text comments held within structured surveys were analysed using a supervised approach.[6 27 28 34 47 48 50 52 54]

There is little evidence in the literature on the statistical properties for the minimum text size needed to perform language analysis, principally because of the difficulty of natural language understanding and the content and context of a text corpus.[6] The studies that reported text size demonstrate that the average character count was around 40 words. The domain of patient feedback from free-text complementing structured surveys appears fixed in its nature, making it attractive data for supervised learning.[31] Just as the domain is fixed, the perspective of a patient feedback document is also fixed[31]: there is limited vocabulary that is useful for commenting about health service, and therefore it is possible to anticipate the meaning of various phrases and automatically classify the comments.[34] Rastegar-Mojarad *et al*[57] also observed that a small (25%) vocabulary set covered a majority (92%) of the content of their patients comments, consistent with a study[60] exploring consumer health vocabulary used by consumers and healthcare professionals. This suggests

that patients use certain vocabulary when expressing their experience within free-text comments.

The overall domain of patient feedback is the healthcare system,[31] and this study revealed the content of reviews tend to focus on a small collection of aspects associated with this as demonstrated by the topics used for text classification in the studies.[24 27 34 48 52] In contrast, the studies[26 32 51 53 55 56] that performed topic modelling, did so on the premise that patient feedback comments contain a multitude of different topics. Topic modelling can be useful in evaluating how close results come to what humans with domain knowledge have determined the topics to be, and if this unsupervised approach finds new topics not identified by humans.[49] LDA was used to extract a number of topics from the free-text reviews as they occur in the data without any prior assumption about what patients care about. The topics identified by six studies that used LDA did not generate any new topics, which is in keeping with the earlier finding that consumer healthcare reporting has limited vocabulary. This finding was supported by Doing-Harris *et al*,[24] who showed that their topic modelling results echo topic classification results, demonstrating that no unexpected topics were found in topic modelling.

Other factors should be taken into account when employing LDA. LDA is mainly based on frequency on co-occurrence of words under similar topics.[51] Topics discovered using LDA may not match the true topics in the data, and short documents, such as free-text comments, may result in poor performance of LDA.[49] In addition to the short comments, studies in this review also demonstrate that majority of the comments on social media tend to be positive, in contrast to the negative reviews which are longer but less frequent. Wagland *et al*[48] found that the content of positive comments was usually much less specific that for negative comments. Therefore an unsupervised approach to short positive reviews may not detect new topics, and the low frequency of negative reviews may not highlight new topics either. To mitigate this, there is a role of using a supervised approach to identify subcategories for negative reviews.[48]

Choice of the number of topics for LDA model also affects the quality of the output.[25 56] If topics are too few, their content gives insight into only very general patterns in the text which are not very useful. Too many topics, on the other hand, make it difficult to find common themes with numerous topics. An LDA topic model with an optimal number of topics should demonstrate meaningful patterns without producing many insignificant topics. The number of topics identified in the studies reviewed[26 32 51 53 55 56] was not consistent and ranged from 6 to 60, demonstrating that deciding on the optimal number is challenging. Performance of the LDA models is affected by semantic coherence (the rate at which topic's most common words tend to occur together in the same reviews) and exclusivity (the rate at which most common terms are exclusive to individual topics). Both measures are useful guidance of which model to choose,[55] however, of the six studies that used LDA, only one study[55] reported LDA performance measures.

Sentiment analysis was commonly conducted using a supervised approach (n=8).[6 24 26 27 34 47 48 54] Even though pre-classified, understanding what the comments both negative and positive are specifically talking about still requires reading through the comments. NLP makes this process efficient by identifying trends in the comment by sentiment. This review identified the most common approach to sentiment classification was to categorise the comment into a single category, that is, positive or negative. However, this implies that there must be polarity associated with a document, which is not always the case. This fails to capture the mixed sentiments or neutral sentiments which could provide useful insights into patient experience. Nawab *et al*[50] demonstrated that splitting the mixed sentiments by sentences revealed distinct sentiments. Therefore, although the percentage of mixed or neutral sentiment is low compared with the overall dataset, analysis of comments within these mixed and neutral sentiment can provide useful information and therefore should not be discarded.

Greaves *et al*[27] and Bahja *et al*[26] used the associated star rating within the NHS Choices data to directly train the sentiment tool. This approach is able to make use of the implicit notion that if a patient says they would recommend a hospital based on star rating, they are then implying a positive sentiment, and conversely if not a negative sentiment, therefore automatically extracting a nominal categorisation. This automated classification removes the need for manual classification and eliminates potential biases of reviewer assignment of comments, but it makes an assumption that star ratings correlate with the sentiment. This is supported by Kowalski,[55] who demonstrated intuitive relationships between topics' meanings and star rating across the analysed NHS Choices dataset. In contrast, Alemi *et al*[34] found that sentiment in comments from RateMDs are not reflected in the overall rating, for example 6% of the patients who gave highest overall rating still included a complaint in their comments, and 33% of patients who gave lowest overall rating included praise. This suggests that the sentiment may not always correlate with the star rating, and therefore researchers need to recognise that the approach used for classification may have implications on validity.

With regard to sentiment analysis of Twitter dataset, Greaves *et al*[28] found no associations when comparing Twitter data to conventional metrics such as patient experience, Hawkins *et al*[52] found no correlation between twitter sentiment and HCAHPS score, suggesting twitter sentiment must be treated cautiously in understanding quality. Therefore, although star ratings can be informative and in line quantitative measures of quality, they may not be sufficiently granular to help evaluate service quality based solely on the star rating without considering the textual content.[53]

Studies in this review demonstrate that NLP and ML have emerged as an important tool for processing

patient experience unstructured free-text data and generating structured output. However, most of the work has been done on extracting information from social media.[6 26–28 32 34 51–58] Healthcare organisations have raised concerns about the accuracy or comments expressed on social media,[61] making policymakers reluctant to endorse narrative information as a legitimate tool. Even though most administrators remove malicious messages manually, anyone can comment on the website and intentionally distort how potential patients evaluate healthcare services. The validity and reliability of NLP is further limited by the fact that most patients do not post reviews online. Kowalski[55] found that healthcare services in England received fewer that 20 reviews over a period of three and a half years. For a limited amount of data, NLP may not be very expedient, and with a smaller number of comments the results may not be as fruitful and there may not be enough raw data to detect a specific pattern.[50] Furthermore, rating posted in social media reviews is not adjusted for user characteristics or medical risk, whereas structured survey scores are patient mix adjusted.[6]

Limitations

We focused on indicators of quality of the included articles rather than assessing the quality of the studies, as relevant formal standards have yet to be established for NLP articles. Due to the heterogenous nature of the studies, and various approaches taken with regard to pre-processing, manual classification and performance of classifiers, it is challenging to make any comparative statements.

CONCLUSION

Studies in this review demonstrate that NLP and ML have emerged as an important tool for processing unstructured free-text patient experience data. Both supervised and unsupervised approaches have their role in language analysis depending on the data source. Supervised learning is time consuming due to the manual coding required, and is beneficial in analysing free-text comments commonly found in structured surveys. As the volume of comments posted on social media continues to rise, manual classification for supervised learning may not be feasible due to time constraints and topic modelling may be a satisfactory approach. To ensure that every patients' voice is heard, healthcare organisations must react and mould their language analysis strategy in line with the various patient feedback platforms.

Acknowledgements We thank Jacqueline Cousins (Library Manager and Liaison Librarian at Imperial College London) for the support improving the composition of the search terms and procedural aspects of the search strategy.

Contributors MK, JS and EM contributed to conception and design of the work. MK and PA contributed to database searching. MK, PA and EM contributed to full-text screening. MK, PA, KF, JS and EM contributed to data analysis and interpretation.

Funding This work is supported by the National Institute for Health Research (NIHR) Imperial Patient Safety Translation Research Centre. Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre.

Disclaimer 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.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

ORCID iD

Mustafa Khanbhai http://orcid.org/0000-0002-4434-1785

REFERENCES

1 Darzi A. High quality care for all: NHS next stage review final report department of health, 2008. Available: www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_085825
2 Coulter AFR, Cornwell J. *Measures of patients' experience in hospital: purpose, methods and uses*. Kings Fund, 2009.
3 Coulter A. What do patients and the public want from primary care? *BMJ* 2005;331:1199–201.
4 Coulter A, Cleary PD. Patients' experiences with hospital care in five countries. *Health Aff* 2001;20:244–52.
5 NHS England. *The friends and family test*. Publication Gateway Ref, 2014.
6 Huppertz JW, Otto P. Predicting HCAHPS scores from hospitals' social media Pages: a sentiment analysis. *Health Care Manage Rev* 2018;43:359–67.
7 Greaves F, Ramirez-Cano D, Millett C, *et al*. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf* 2013;22:251–5.
8 López A, Detz A, Ratanawongsa N, *et al*. What patients say about their doctors online: a qualitative content analysis. *J Gen Intern Med* 2012;27:685–92.
9 Trigg L. Patients' opinions of health care providers for supporting choice and quality improvement. *J Health Serv Res Policy* 2011;16:102–7.
10 Cognetta-Rieke C, Guney S. Analytical insights from patient narratives: the next step for better patient experience. *J Patient Exp* 2014;1:20–2.
11 Robert G, Cornwell J. Rethinking policy approaches to measuring and improving patient experience. *J Health Serv Res Policy* 2013;18:67–9.
12 Ipsos-MORI. *Real time patient feedback: information patients need and value, research report prepared for West Midlands strategic health authority*, 2008.
13 Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. *Int J Med Inform* 2018;114:57–65.
14 Yim W-W, Yetisgen M, Harris WP, *et al*. Natural language processing in oncology: a review. *JAMA Oncol* 2016;2:797–804.
15 Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. *Methods* 2015;74:97–106.
16 Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018;77:34–49.
17 Gibbons C, Richards S, Valderas JM, *et al*. Supervised machine learning algorithms can classify Open-Text feedback of doctor performance with Human-Level accuracy. *J Med Internet Res* 2017;19:e65.
18 Chowdhury GG, Cronin B. Natural language processing. *Ann Rev Info Sci Tech* 2002;37:51–89.

19 Hotho A, Nurnberger A, Paass G. A brief survey of text mining. *Ldv Forum* 2005;20:19–62.
20 Feldman R, Sanger J. *The text mining Handbook: advanced approaches in analyzing unstructures data*. Cambridge University Press, 2007.
21 Collingwood L, Wilkerson J. Tradeoffs in accuracy and efficiency in supervised learning methods. *J Inf Technol* 2012;9:298–318.
22 Alloghani M, Al-Jumeily D, Mustafina J. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: *Supervised and unsupervised learning for data science*. Springer, Cham, 2020.
23 Kotsiantis SB. Supervised machine learning: a review of classification techniques. *Informatica* 2007;31:249–68.
24 Doing-Harris K, Mowery DL, Daniels C, *et al*. Understanding patient satisfaction with received healthcare services: a natural language processing approach. *AMIA Annu Symp Proc* 2016;2016:524–33.
25 Blei DNA, Jordan M. Latent Dirichlet allocation. *J Mach Learn Res* 2003;3:993–1022.
26 Bahja MLM. Identifying patient experience from online resources via sentiment analysis and topic modelling. *Association for Computing Machinery* 2016;6.
27 Greaves F, Ramirez-Cano D, Millett C, *et al*. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J Med Internet Res* 2013;15:e239.
28 Greaves F, Laverty AA, Cano DR, *et al*. Tweets about Hospital quality: a mixed methods study. *BMJ Qual Saf* 2014;23:838–46.
29 Liu B. *Sentiment analysis and opinion mining*. San Rafael, California: Morgan & Claypool Publishers, 2012: 5. 1–167.
30 Pang B, Lee L. Opinion mining and sentiment analysis. *FNT in Information Retrieval* 2008;2:1–135.
31 Smith P. *Sentiment analysis of patient feedback*. University of Birmingham, 2015.
32 Wallace BC, Paul MJ, Sarkar U, *et al*. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *J Am Med Inform Assoc* 2014;21:1098–103.
33 Gohil S, Vuik S, Darzi A. Sentiment analysis of health care Tweets: review of the methods used. *JMIR Public Health Surveill* 2018;4:e43.
34 Alemi F, Torii M, Clementz L, *et al*. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Qual Manag Health Care* 2012;21:9–19.
35 Sheard L, Marsh C, O'Hara J, *et al*. The Patient Feedback Response Framework - Understanding why UK hospital staff find it difficult to make improvements based on patient feedback: A qualitative study. *Soc Sci Med* 2017;178:19–27.
36 The Power of Information. *Putting all of US in control of the health and care information we need*. London: Department of Health, 2012.
37 Griffiths A, Leaver MP. Wisdom of patients: predicting the quality of care using aggregated patient feedback. *BMJ Qual Saf* 2018;27:110–8.
38 Gibbons C, Greaves F. Lending a hand: could machine learning help hospital staff make better use of patient feedback? *BMJ Qual Saf* 2018;27:93–5.
39 Rozenblum R, Greaves F, Bates DW. The role of social media around patient experience and engagement. *BMJ Qual Saf* 2017;26:845–8.
40 Department of Health. *What matters: a guide to using patient feedback to transform services*, 2009.
41 Francis R. *Report of the mid Staffordshire NHS Foundation trust public inquiry*, 2013.
42 Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
43 Pons E, Braun LMM, Hunink MGM, *et al*. Natural language processing in radiology: a systematic review. *Radiology* 2016;279:329–43.
44 Mishra R, Bian J, Fiszman M, *et al*. Text summarization in the biomedical domain: a systematic review of recent research. *J Biomed Inform* 2014;52:457–67.
45 Dreisbach C, Koleck TA, Bourne PE, *et al*. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* 2019;125:37–46.
46 Koleck TA, Dreisbach C, Bourne PE, *et al*. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019;26:364–79.
47 Menendez ME, Shaker J, Lawler SM, *et al*. Negative Patient-Experience comments after total shoulder arthroplasty. *J Bone Joint Surg Am* 2019;101:330–7.
48 Wagland R, Recio-Saucedo A, Simon M, *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;25:604–14.
49 Rivas C, Tkacz D, Antao L, *et al*. *Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study. health services and delivery research*. Southampton (UK), 2019.
50 Nawab K, Ramsey G, Schreiber R. Natural language processing to extract meaningful information from patient experience feedback. *Appl Clin Inform* 2020;11:242–52.
51 Hao H, Zhang K. The voice of Chinese health consumers: a text mining approach to web-based physician reviews. *J Med Internet Res* 2016;18:e108.
52 Hawkins JB, Brownstein JS, Tuli G, *et al*. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf* 2016;25:404–13.
53 James TL, Villacis Calderon ED, Cook DF. Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Syst Appl* 2017;71:479–92.
54 Jimenez-Zafra SM M-VM, Maks I, Izquierdo R. Analysis of patient satisfaction in Dutch and Spanish online reviews 2017;58:101–8.
55 Kowalski R. Patients' written reviews as a resource for public healthcare management in England. *Procedia Comput Sci* 2017;113:545–50.
56 Ranard BL, Werner RM, Antanavicius T, *et al*. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Aff* 2016;35:697–705.
57 Rastegar-Mojarad M, Ye Z, Wall D, *et al*. Collecting and analyzing patient experiences of health care from social media. *JMIR Res Protoc* 2015;4:e78.
58 Plaza-del-Arco F M-VT, Jimenez-Zafra SM, Molina-Gonzalez D. COPOS: corpus of patient opinions in Spanish. Application of sentiment analysis techniques. *Procesamiento del Lenguaje Natural* 2016;57:83–90.
59 Haddi E, Liu X, Shi Y, Xiaohui L, Yong S. The role of text pre-processing in sentiment analysis. *Procedia Comput Sci* 2013;17:26–32.
60 Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc* 2006;13:24–9.
61 McCartney M. Will doctor rating sites improve the quality of care? no. *BMJ* 2009;338:b1033.