

The use of natural language processing to interpret unstructured patient feedback on health services: A scoping review

Ali Feizollah, Chiu-Yi Lin, Lucy O'malley, Wendy Thompson, Stefan Listl, Matthew Byrne

Submitted to: Journal of Medical Internet Research
on: February 20, 2025

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

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	39
.....	39
Multimedia Appendixes	40
Multimedia Appendix 1.....	40
CONSORT (or other) checklists.....	41
CONSORT (or other) checklist 0.....	41

The use of natural language processing to interpret unstructured patient feedback on health services: A scoping review

Ali Feizollah¹; Chiu-Yi Lin¹; Lucy O'malley¹; Wendy Thompson¹; Stefan Listl²; Matthew Byrne¹

¹ University of Manchester Manchester GB

² University Hospital Heidelberg Heidelberg DE

Corresponding Author:

Ali Feizollah

University of Manchester
Oxford Road
Manchester
GB

Abstract

Background: Unstructured patient feedback (UPF) allows patients to freely express their experiences without the constraints of predefined questions. The proliferation of online healthcare rating websites has created a vast source of UPF. Natural language processing (NLP) techniques, particularly sentiment analysis and topic modelling, are increasingly being used to analyse UPF in healthcare settings, however the scope and clinical relevance of these technologies is unclear.

Objective: This scoping review investigates how NLP techniques are being used to interpret UPF, with focus on the healthcare settings in which this is used, the purposes for using these technologies, and any impacts reported on clinical practice.

Methods: Searches of the MEDLINE, EMBASE, CINAHL, Cochrane Database of Reviews, and Google Scholar were conducted in February 2024. No date limits were applied. English language studies that used NLP techniques on UPF that pertained to an identifiable health care setting or provider were included. Data extraction focused on the healthcare setting, NLP methods used, and applications of these techniques.

Results: 52 studies were included. NLP was most commonly applied to UPF from secondary care settings (n=33) with fewer in primary (n=10) or community (n=5) care. Three NLP techniques were identified in the included studies: sentiment analysis (n=32), topic modelling (n=15) and text classification (n=7). Sentiment analysis was applied to explore associations between patient sentiment and healthcare provider characteristics, track emotional responses over time, and identify areas for improvement in healthcare delivery. Topic modelling, primarily using Latent Dirichlet Allocation (LDA) algorithm, was employed to uncover latent themes in patient feedback, compare patient experiences across different healthcare settings, and track changes in patient concerns over time. Text classification was used to categorize patient feedback into predefined topics. The association between NLP-derived insights and traditional healthcare quality metrics was limited, with few studies describing concrete clinical impacts resulting from their analyses.

Conclusions: NLP has been applied to UPF across a number of contexts, primarily to identify features of health services or professionals that support good patient experience. The growth of research publications demonstrates an academic interest in these technologies, but there is little evidence these approaches are being employed in clinical settings. Future research is required to assess how NLP may capture the nuance of healthcare interactions, align with existing quality metrics and how it may be used to influence clinician behaviour

(JMIR Preprints 20/02/2025:72853)

DOI: <https://doi.org/10.2196/preprints.72853>

Preprint Settings

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

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

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

Only make the preprint title and abstract visible.

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

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

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

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

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

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/>



Original Manuscript

Paper type: original paper

The use of natural language processing to interpret unstructured patient feedback on health services: A scoping review

Ali Feizollah^{1*}, Chiu-Yi Lin¹, Lucy O'malley¹, Wendy Thomson¹, Stefan Listl², Matthew Byrne¹

¹*Division of Dentistry, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK*

² *Section for Oral Health, Heidelberg Institute of Global Health, Heidelberg University Hospital (UKHD), Heidelberg, Germany*

Corresponding Author: Ali Feizollah
ali.feizollah@manchester.ac.uk

Abstract

Background:

Unstructured patient feedback (UPF) allows patients to freely express their experiences without the constraints of predefined questions. The proliferation of online healthcare rating websites has created a vast source of UPF. Natural language processing (NLP) techniques, particularly sentiment analysis and topic modelling, are increasingly being used to analyse UPF in healthcare settings, however the scope and clinical relevance of these technologies is unclear.

Objectives:

This scoping review investigates how NLP techniques are being used to interpret UPF, with focus on the healthcare settings in which this is used, the purposes for using these technologies, and any impacts reported on clinical practice.

Methods:

Searches of the MEDLINE, EMBASE, CINAHL, Cochrane Database of Reviews, and Google Scholar were conducted in February 2024. No date limits were applied. Eligibility criteria included English language studies that used NLP techniques on UPF that pertained to an identifiable health care setting or providers. Studies were excluded if human actors solely performed coding or if NLP was applied to structured feedback or non-patient-generated content. Data was extracted and narratively synthesized regarding healthcare settings, NLP methods, and clinical applications.

Results:

From 4,017 records, 52 studies met inclusion criteria. NLP was most commonly applied to UPF from secondary care settings (n=33) with fewer in primary (n=10) or community (n=5) care. Three NLP techniques were identified in the included studies: sentiment analysis (n=32), topic modelling (n=15) and text classification (n=7). Sentiment analysis was applied to explore associations between patient sentiment and healthcare provider characteristics, track emotional responses over time, and identify areas for improvement in healthcare delivery. Topic modelling, primarily using Latent Dirichlet Allocation (LDA) algorithm, was employed to uncover latent themes in patient feedback, compare patient experiences across different healthcare settings, and track changes in patient concerns over time. Text classification was used to categorize patient feedback into predefined topics. The association between NLP-derived insights and traditional healthcare quality metrics was limited, with few studies describing concrete clinical impacts resulting from their analyses.

Conclusions:

NLP has been applied to UPF across a number of contexts, primarily to identify features of health services or professionals that support good patient experience. The growth of research publications demonstrates an academic interest in these technologies, but there is little evidence these approaches are being employed in clinical settings. Future research is required to assess how NLP may capture the nuance of healthcare interactions, align with existing quality metrics and how it may be used to influence clinician behaviour.

Keywords:

Natural Language Processing; Patient Reported Outcome Measures; Patient Satisfaction; Data Mining; Quality of Health Care; Social Media; Machine Learning; Quality Improvement; Medical Informatics; Patient Generated Health Data

Introduction

Patient experience is frequently used as an indicator of quality in health care systems [1]. Health care services that provide their patients with good experiences are more likely to retain patients; retention of patients both improves patient outcomes [2] and assures the business viability of health services [3]. As such, health services make efforts to measure patient experience as part of quality assurance schemes.

To capture patient experiences, healthcare providers use various feedback mechanisms. Structured Patient Reported Experience Measures (PREMs) are commonly used to assess service quality [4], but are limited to the narrow concept that they intend to measure. In contrast, Unstructured Patient feedback (UPF) allows patients to describe their experiences of care without the restriction of preselected questions [5]. Health providers may use UPF in local quality improvement efforts, using 'suggestion boxes', testimonials, free text feedback forms and written patient complaints. By freely expressing their experiences, patients may identify issues that relate to high- and low-quality care that may not be measured in other ways [6, 7].

Whilst UPF provides rich insights, traditional methods of collecting, collating, and interpreting these data at large scales are time-consuming and resource intensive [8]. This challenge has been amplified by the proliferation of online healthcare rating sites, which have become ubiquitous sources of patient experience and satisfaction data [9, 10]. Unlike local feedback mechanisms, these platforms are not usually coordinated by the health service being reviewed, creating a vast, decentralized source of patient feedback. Studies associating patient experiences reported in online reviews and physician performance have demonstrated both poor [11, 12], and good [13] correlation between patient experience and traditional performance metrics.

The proliferation of online reviews means that this is now a large data source that crosses a number of areas of healthcare [10, 14]. The rate and scale at which new reviews are generated means that it would be infeasible for assessments of the state of different health to be made through conventional methodologies such as thematic analysis. As these data sources continue to expand, healthcare systems require more efficient approaches to extract meaningful insights from patient feedback.

Natural language processing (NLP) offers a solution by using machine learning and artificial intelligence algorithms to interpret text data [15] in a time and resource efficient manner [16]. Beyond efficiency, NLP offers several advantages over conventional thematic analysis: it reduces human coding bias [17], enables reproducible analysis at scale [18], allows detection of subtle patterns that might not be apparent to human analysts [19], and permits longitudinal analysis of patient feedback over time to detect evolving trends [20]. NLP includes sentiment analysis, which assesses whether a text has an overall positive or negative sentiment to show people's opinions, attitudes and emotions [21], topic modelling which determines the frequency and association of words within texts to develop topics of interest [22, 23] and topic classification, where texts are classified to preselected topics.

Despite the growing application of NLP to patient feedback, there remains a significant gap in understanding how these methods are applied across different health settings. The extent to which NLP analysis of UPF is being used to drive quality improvement across different healthcare settings

is currently unclear, limiting its potential impact on patient care. While NLP applications to patient experience has been reported in contexts such as hospital care [24, 25], general medical practice [26], and dentistry [27, 28], the comprehensive landscape remains unmapped. A systematic review by Khanbhai et al was performed in 2021 [29] exploring the use of machine learning and NLP in patient experience feedback. The search conducted in this paper occurred in 2020, with 15/19 included papers identified in the years 2015-2020. Preliminary literature searching has demonstrated a proliferation in studies exploring the use of NLP since the completion of this review, suggesting a more current and comprehensive assessment.

The aim of this scoping review is to investigate how NLP techniques are currently being used to interpret unstructured patient feedback across different health services. Objectives of this review were to identify the settings in which NLP been used to interpret patient feedback, the purposes using NLP and to understand if the interpretation of patient feedback with NLP been used to inform any changes in clinical practice or policy.

Methods

Eligibility criteria

This scoping review follows the Joanna Briggs Institute (JBI) methodology for scoping reviews [30]. As this review is looking at patient feedback provided across all types of healthcare, no specific population was defined.

For inclusion, sources must have explored of the use of NLP on UPF of a health service or provider. NLP was defined as computer based algorithmic assessment of text, with or without training of probabilistic models. UPF was defined as any unstructured text or 'free text', written by a user of a health service to relate their outcomes or experience of that health service. The contexts explored in this review were health services, including, but not limited to medicine, dentistry, physiotherapy, pharmacy, and ophthalmology. Papers that explored UPF that was attributable to a particular health service or provider were included. Primary sources of UPF in these papers may include in-clinic collection on comment cards, hospital websites online service rating platforms or social media platforms.

Sources were excluded if coding and interpretation of UPF was performed by a human actor only. Studies using NLP on structured feedback, electronic health records, and chatbots were excluded because they represent fundamentally different data types and clinical applications. Structured feedback contains predefined response categories with limited expressive range, electronic health records contain clinician rather than patient language, and chatbot interactions represent dialogues rather than retrospective experiences.

Sources from the peer reviewed and grey literature were considered. This included both experimental and quasi-experimental study designs including randomized controlled trials, non-randomized controlled trials, before and after studies and interrupted time-series studies. In addition, analytical observational studies including prospective and retrospective cohort studies, case-control studies and analytical cross-sectional studies were considered for inclusion. Descriptive

observational study designs including case series, individual case reports and descriptive cross-sectional studies were considered for inclusion. In addition, systematic reviews that meet the inclusion criteria will also be considered, depending on the research question as a source of further articles for inclusion of further articles. Conference abstracts were also considered.

Search strategy

An initial limited search of MEDLINE and EMBASE was undertaken to identify articles. Key words contained in the titles and abstracts of relevant articles, and the index terms used to describe the articles were used to develop a full search strategy for MEDLINE, EMBASE, CINAHL, and Cochrane Database of Reviews (Multimedia Appendix 1). Reference lists of all included sources of evidence were screened for additional studies. Where studies were not available, authors were contacted directly to attempt to attain copies. To try and identify further peer reviewed literature and grey literature Google Scholar and Google were searched in advanced search function using keywords of "patients," "patient feedback," "health care," "NLP," "topic modelling", "topic classification" and "Sentiment analysis." Initial searches for grey literature identified limited information, primarily comprising opinion pieces on the promise of NLP rather than the impact of NLP on health services. In interest of time and resources for the review, peer reviewed sources were prioritised.

Study/Source of Evidence selection

All identified sources were collated in Endnote 20 (Clarivate Analytics, PA, USA) and duplicates were removed. Pilot screening was performed by two independent reviewers on a sample of 20 papers to ensure consistent application of the inclusion and exclusion criteria. Following this, all titles and abstracts were screened by single authors using Rayyan [31]. Full texts of shortlisted papers were assessed in detail against the inclusion and exclusion criteria by two independent reviewers (AF and CL). Disagreements were resolved through regular meetings, with a third reviewer (MB) acting as a tie-breaker. Reasons for exclusion of sources of evidence at full text that did not meet the inclusion criteria were recorded and are reported in the PRISMA flow diagram. While formal critical appraisal was not conducted as per scoping review methodology, we assessed the reliability of included studies through research team discussions during data extraction meetings to identify methodological concerns. We considered the peer-review process of the journals in which studies were published as a quality filter. Additionally, we considered journal accessibility to ensure our review captured studies that would be readily available to healthcare practitioners and researchers interested in implementing NLP approaches.

Data Extraction

Data was independently extracted by three reviewers (AF, CL, and MB) using a data extraction spreadsheet developed by the reviewers (Multimedia Appendix 2). Heterogeneity of methods, contexts and reporting meant that quantitative comparison was not possible. Qualitative assessment and narrative summaries of reviews were prepared in the extraction tool. The research team conducted an internal review of the extraction process through regular meetings to ensure consistency. Following initial extraction, further elements to the form were added, to record any performance metrics associated with the NLP methods. As a scoping review, critical appraisal of individual pieces of evidence was not carried out.

Results

Search results

Five databases were searched in February 2024, yielding 4017 records: MEDLINE (n=675), EMBASE (n=2,286), CINAHL (n=1,024), Cochrane Database of Reviews (n=3), and Google Scholar (n=29). After removing duplicates, 3,433 were screened by two independent reviewers, resulting in 3,052 exclusions. An additional 283 records were marked as "maybe," and 64 records had conflicts that required discussion. Following discussions with a third reviewer, 92 records were included while 3,341 were excluded.

Full-text screening of these 92 records resulted in the inclusion of 52 studies and the exclusion of 40 studies based on the predefined inclusion and exclusion criteria (Figure 1). The summary of the included studies is reported in Multimedia Appendix 3.

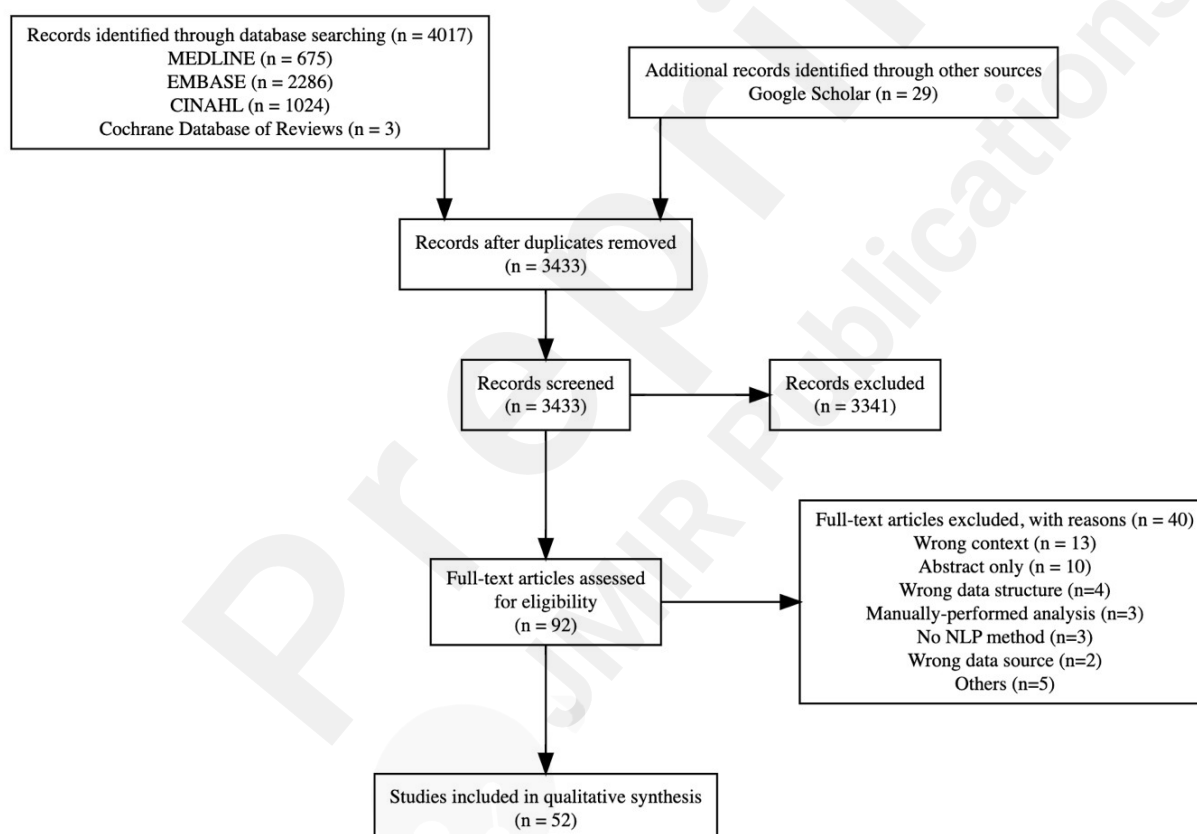


Figure 1. PRISMA flow diagram

Healthcare settings in which NLP has been used to assess UPF

Seven research papers were associated with health services in English-speaking nations (United States = 3, United Kingdom = 4). Ten studies were carried out in non-English-speaking nations (Iran =3, India, China =2, South Korea, Spain, Netherlands = 1). Three broad categories of NLP were used: sentiment analysis, topic modelling, or text classification. The distribution of reviewed papers based on their healthcare setting (primary, secondary, or community care) and the specific NLP method used is presented in Table 1. Note that the total count in the table is greater than the number of

reviewed papers, as some papers use more than one method in their analysis.

Table 1. Distribution of NLP Review Papers by Healthcare Setting and NLP Method (note some studies used multiple methods)

	Sentiment analysis	Topic modelling	Text classification
Primary care	7	1	2
Secondary care	21	9	3
Community	2	3	0
Unspecified	2	2	2
Total	32	15	7

Applications of NLP

Sentiment Analysis

Sentiment analysis was used in 32 of the 52 studies considered. VADER [32], a 'stock' open source sentiment analysis tool, was used in 11 studies [21, 33-42]. Basic machine learning algorithms were used in 8 of the reviewed papers for sentiment analysis, including support vector Machine learning [43-47], Naïve Bayes [45-48] and Decision Tree [46, 49]. Advanced neural networks were used in 3 of the reviewed papers and included the Keras library [16], Recurrent neural networks [50] and convolutional neural networks [51]. Third-party paid sentiment analysis services used included Crystalfeel [52, 53], Baidu [54], Tencent [55], IBM Watson [56] and the Press Ganey Associates' NLP tools for surveys and feedback forms [57].

A number of studies used sentiment analysis to explore the associations between patient sentiment and demographic factors of clinicians, such as their age and their geographical location. Studies examining online reviews of otolaryngologists [33], spine surgeons [36], neurosurgeons [39], urologists [41] and psychiatrists [40] found that younger practitioners received higher sentiment scores. Location was shown to be a significant factor in sentiment scores of neurosurgeons [39], psychiatrist [40] and otolaryngologists [33]. Several studies in this review demonstrate a strong correlation between sentiment scores derived from patient reviews and the star ratings given to healthcare providers [16, 44, 51, 56].

Word frequency analysis was often used alongside sentiment to characterise patient sentiments their interactions with healthcare. By analysing the language used in online reviews and comments, researchers can uncover common themes, concerns, and areas for improvement in healthcare delivery [33]. For example, Park et al. found that, for psychiatrists, positive reviews mentioned "time" and "caring", whereas negative reviews mentioned "medication" [40]. Pain management was identified as a significant driver of patient satisfaction across various surgical specialties, including hand surgery [34], scoliosis surgery [35], and spine surgery [36]. Gour and Kumari used fuzzy sentiment analysis and word frequency analysis to identify trust and fear as components of positive and negative reviews [58]. Nawab et al. combined sentiment analysis with word frequency to associate hospital room conditions and discharge processes as markers of quality [16].

Sentiment analysis was also used to track patients' emotional responses to healthcare services over time. Shah et al. used aspect-based sentiment analysis to track changes in patient sentiment and

topics of concern expressed in online reviews over the course of the COVID-19 pandemic showing that fear, anger, and sadness in the early stages of the pandemic gradually shifted towards more positive sentiments [53]. Similarly, Li et al. used sentiment and word frequency analysis to investigate changes in doctor-patient relationships during COVID-19, noting a shift in the focus of negative comments from personal attitudes to administrative issues [54]. Further applications of sentiment analysis include assessment of transitions of care and continuity [59]. Hu et al. monitored public perceptions on healthcare services using social media data in China [55]. The findings indicated that the doctor-patient relationship category had the highest proportion of negative contents, followed by service efficiency, and nursing service.

Topic Modelling

Topic modelling methods were applied in 15 of the 52 studies. Topic modelling is an unsupervised machine learning technique that can scan large volumes of text to identify latent topics based on word co-occurrence [60]. Yazdani et al. used topic modelling on patient feedback of hospitalized cancer identifying dissatisfaction with appointment booking services and positive experiences with staff and chemotherapy [61]. Stokes et al. analysed narrative themes in online reviews of mental health facilities, identifying that caring staff and non-pharmacologic treatment modalities were positively correlated with high ratings, while issues like safety and abuse and poor communication were linked to negative reviews [62]. Agarwal et al. compared online ratings of patient experiences between emergency departments and urgent care centres, finding comfort, professionalism, and staff interactions being key themes in 5-star reviews for both types of facilities [59].

Lin et al. applied topic modelling to online reviews of dental care, finding that higher ratings were associated with female dentists, dentists at a younger age, and those whose patients experienced a short wait time. They also identified several topics that corresponded to CAHPS measures [63], including discomfort (e.g., painful/painless root canal or deep cleaning), and ethics (e.g., high-pressure sales, and unnecessary dental work) [27].

Latent Dirichlet Allocation (LDA) is a type of topic modelling that assumes that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics [60]. Ranard et al. used LDA to identify topics in Yelp reviews of hospitals, demonstrating that reviews covered more topics than the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey [64], including the cost of hospitals visits, insurance, billing, and the quality of nursing and staff [7]. LDA analysis of pharmacy Yelp reviews by Lester and Chui revealed four key topics: prescription wait times, staff helpfulness, store environment, and medication filling issues. Pearson correlations showed that wait times and filling issues negatively correlated with ratings, while staff helpfulness and store environment showed positive correlations [65].

Extensions and modifications to LDA were used in a number of papers: Shah et al. developed Topic Coherence-Based LDA (TCLDA), to improve the interpretability of topics by measuring the semantic similarity between high-scoring words in a topic. TCLDA was used to identify emerging and fading topics in patient online reviews during the early wave of the COVID-19 pandemic showing an increased focus on treatment experiences, policy implementation, and mental health developed over time [52]. In a further study, Shah et al. used dynamic topic modelling to investigate the dynamics of public concerns and sentiments expressed in 2018, 2019, and 2020 [66]; This showed

that topics shifted from general healthcare issues to pandemic-specific concerns like virus transmission, travel restrictions, and government countermeasures. Sentiments initially became more negative, with anger as the dominant emotion [53].

Gibbs Sampling Dirichlet Mixture Model (GSDMM), a modified version of LDA that assumes that each document has only one topic, making it more efficient for dealing with short texts [67]. Serrano-Guerrero et al. used GSDMM to group sentences from patient opinions and identify the most frequent topics like coordination, scheduling appointments, order queue, community support, etc. related to nurses and doctors in different healthcare categories (high-risk disease, low-risk disease, and infectious disease) [21].

Non-Negative Matrix Factorization (NMF) factorizes a matrix into two non-negative matrices [68] for dimensionality reduction and feature extraction and has been reported to provide more interpretable topics than LDA [69]. Langerhuizen et al. used NMF to identify the 50 most frequently occurring three-word combinations (for example poor bedside manner, office staff rude, waited lobby hour, etc.) in online reviews of orthopaedic surgeons and office, identifying themes such of logistics, care and compassion, trust, recommendation, and customer service as important elements of quality [56]. Tones of joy and confidence were associated with higher ratings. Sadness and tentative tones were associated with lower ratings.

Text Classification

Seven studies conducted text classification analysis. Text classification uses machine learning to assign predefined categories or labels to textual data. This process involves training a model on a labelled dataset of texts that provides the 'ground truth' from which the model can learn to predict the labels of new, unseen text [70]. This review identified two instances where unstructured patient feedback was used to train NLP text classifiers.

Khanbhai et al. applied supervised learning algorithms to categorize patient feedback [43]. This involves training models on labelled datasets to classify new, unlabelled data accurately. The study also utilizes topic classification tools like the KoNstanz Information Miner (KNIME) platform for qualitative content analysis [71]. This approach helps in systematically categorizing patient concerns into distinct topics, providing a structured overview of patient feedback.

Similarly, He et al. integrate both supervised and unsupervised machine learning approaches. Initially, a set of reviews is manually coded to identify major themes, a process known as qualitative coding [72]. This coded data then serves as the training set for supervised machine learning algorithms, enabling the classifiers to generalize the identified themes across the entire dataset. Moreover, the study employs unsupervised learning techniques, such as word clustering using the k-means algorithm, to identify fine-grained aspects of patient concerns.

Changes to clinical practices reported

Whilst almost all of the studies examined stated the potential of NLP for UPF to inform clinical practices and changes in behaviours, few concrete changes to clinical practice were reported. Parikh et al. describe how NLP is used to analyse patient feedback from MRI scans to identify potential care issues, allowing appropriate clinical teams to be notified and intervene before problems escalate

[73]. Nawab et al. demonstrated how NLP of patient experience comments allowed for quick identification of negative markers of patient experience within a hospital setting, identifying climate control and temperature of waiting rooms as a factor that could be easily modified [16]. Menendez et al. used NLP of negative reviews to identify the main sources of complaints in an orthopaedic hospital, showing that improvements in the quality of patient rooms would likely yield the greatest improvement in patient experience [57]. Khanbhai et al. describe how the use of NLP of free text comment of friend and family test results in 4 hospital trusts in the UK can be used to facilitate rapid action on feedback; while no specific examples of actions implemented were given, the benefit of shifting resource from manual analysis of reviews to implementation of quality improvement actions was highlighted [43]. A similar benefit of time saving of human evaluators of patient feedback was also reported by Khaleghparast [49].

Cammel et al. used sentiment analysis, topic modelling and prioritization factorization to develop top 5 rankings for areas needing improvement and ongoing monitoring within different hospital environments, translating their findings into actionable priorities for hospital improvement initiatives [74].

Alemi and Jasper used NLP alongside traditional summarisation of text to allow managers departments to assess the quality of care across domains related to the CAHPs survey. This was intended to allow for targeted improving of quality improvement activities in different clinical contexts within a hospital [75].

Discussion

Whilst NLP has been routinely used in business and the service industry to analyse customer reviews for service improvement [38, 76], healthcare has been slower to adopt these technologies. This review demonstrates a growing application of NLP to patient feedback in healthcare but a significant gap in knowledge of how best to translate these findings into actions that will improve a patient experience and outcomes. Our scoping review significantly updates and builds on the work of Khanbhai et al. [29], whose 2021 systematic review analysed 19 studies on NLP and machine learning applications for patient experience feedback published until December 2019. Although Khanbhai's review offered valuable insights into this emerging field, our review reveals rapid growth, increasing the number of studies from 19 to 52 within a short period and reflecting heightened academic interest in NLP for patient feedback. Despite this progress, our review confirms that a research-to-practice gap remains, underscoring the need for future work to focus on practical implementation rather than solely on technical feasibility.

Sentiment analysis emerged as the most commonly used form of NLP for UPF in this review. When used alongside word frequency analysis or topic modelling, sentiment analysis may indicate reasons for satisfaction or dissatisfaction with care. The technology has demonstrated practical value in several contexts, including identifying environmental factors affecting patient experience [16], or areas of patient dissatisfaction to inform potential changes to practice [16]. These factors may not have been picked up in traditional measures of patient experience or health care quality, that are limited by the content of pre-conceived questions of researchers or clinicians. However, this review demonstrates a proliferation of studies using simple sentiment analysis tools like VADER to relate patient satisfaction to demographic factors of clinicians such as their age or gender. Such simplistic analyses do not offer much benefit in terms of assessing the quality of clinicians' decision making or

clinical skill; greater interrogation of underlying factors is indicated. Overreliance on simple tools for sentiment analysis may not capture the complexity of healthcare experiences. Many interactions with health services are likely to cause degree of physical or mental discomfort, even if overall outcomes and experiences are good, leading patients to express mixed emotions about their care. More sophisticated approaches using advanced neural networks show promise in capturing these nuances, as demonstrated by Gui and He [51], whose convolutional neural networks outperformed traditional methods in analysing sentiment subtleties in healthcare reviews. Nawab et al. [16] used neural network approaches effectively to identify specific aspects of patient experience. Recent developments in transformer-based models such as BERT [77] and large language models [78] offer potential improvements through their ability to better understand healthcare-specific terminology and context, though their application to patient feedback analysis remains limited in the current literature. Future developments for sentiment analysis should focus on healthcare-specific models that account for the unique context of medical experiences, integration with existing quality improvement frameworks, and validation studies comparing automated sentiment analysis against traditional patient experience measures.

Topic modelling emerged as a practical tool for healthcare administrators to efficiently process large volumes of patient feedback, serving primarily administrative and public relations purposes rather than driving clinical improvements. Studies demonstrate its utility in rapidly identifying operational issues like appointment booking problems in cancer care [61], service quality in mental health facilities [62], and comparing emergency and urgent care services [59]. LDA can reveal insights and latent themes in UPF that would not be identified in traditional surveys, such as billing concerns and wait times [7, 65]. There is limited evidence of these insights translating into concrete clinical or service changes. This suggests a gap between the potential of topic modelling for healthcare quality improvement and its current practical applications in healthcare settings. If topic modelling were to be used more widely for quality improvement purposes, the limitations of this as a technology and the decisions made by researchers in developing models must be appropriately recorded and reported. As topical modelling is an unsupervised machine learning technique, it requires a process of trial and error to yield meaningful results. Researchers must make subjective judgments about the number of topics to extract and interpret the resulting topic clusters, which introduces a potential for bias. The manual labelling of topics identified through LDA, as demonstrated in Lin et al. [23] and Stokes et al. [58], inevitably involves some degree of subjective decision-making. Transparent reporting of the philosophical positions of the researchers labelling and interpreting the topics through reflexivity statements, reinterpretation of data by different research groups, or the involvement of patient and practitioner stakeholder groups in the development of models may help to improve the validity of these techniques in healthcare research.

When considering the effectiveness of different NLP approaches for analysing UPF, it is important to note that performance varies significantly based on multiple factors including data characteristics (volume, length of comments, language complexity), healthcare context, pre-processing techniques employed, and the specific objectives of the analysis. Methods that demonstrate high performance in one setting may not necessarily translate to others with the same effectiveness. Rather than identifying a single 'best' approach, healthcare organizations should consider their specific requirements, available resources, and the nature of their patient feedback data when selecting appropriate NLP methods. This context-dependent performance highlights the need for careful method selection and evaluation when implementing NLP solutions for patient feedback analysis.

A significant issue is the limited association between NLP-derived insights and traditional concepts of healthcare quality. Many established quality metrics in healthcare are based on clinical outcomes, process measures, and standardized patient surveys like HCAHPS and CAHPS. Several studies aimed to demonstrate validity of their models in assessing quality and performance of clinicians by demonstrating correlations between sentiment scores of unstructured and conventional star ratings used to rate an experience as a whole [16, 44, 51, 56]; the relationship to broader healthcare quality metrics remains underexplored. UPF analysed by NLP techniques is a rich source of in-patient perspectives and captures themes that extend beyond traditional survey measures. For example Ranard et al.'s analysis of hospital Yelp reviews revealed that both topics contained within the HCAHPS and those not, including cost of hospital visits, insurance, and billing issues could be identified through NLP of UPF [7]. Similarly, Lin et al.'s analysis of dental care reviews identified additional quality indicators such as discomfort during procedures and ethical concerns about unnecessary treatments [27]. This broader scope of patient feedback, while valuable, may not easily align with traditional metrics. This misalignment between NLP-derived insights and established quality frameworks can lead to scepticism among healthcare professionals and policymakers about the validity and utility of these approaches, despite their potential to capture important aspects of patient experience that traditional metrics might miss.

The scarcity of papers describing tangible clinical outcomes resulting from NLP analyses suggest academic rather translational impact have been the driving force behind research so far. This gap between academic research and clinical practice raises questions about the practical utility of these approaches in improving healthcare delivery. The research community should now look towards implementation science methods to assess how best to use NLP to enact real-world change. In order for feedback interventions to be effective and affect changes in clinical behaviours, they must be targeted, relevant and deemed as trustworthy and valid by clinicians [79]. Greater understanding of the perceived validity of insights developed from NLP must be gained prior to its widespread adoption. Study designs for NLP research should demonstrate clear pathways to clinical application and include measures of clinical impact. Further collaboration between NLP researchers and healthcare providers should be encouraged to ensure research questions address real-world clinical needs. Where clinical impacts are described, it is important to consider whether these occur at the policymaker level, institutional level, or individual clinician level.

Limitations

This scoping review has several limitations that should be considered when interpreting its findings. Firstly, the review was limited to English-language publications, potentially excluding relevant studies published in other languages and introducing a language bias. The rapid evolution of NLP technologies means that some of the most recent advancements may not be fully represented in the published literature included in this review. The heterogeneity of NLP techniques and healthcare settings made it challenging to draw direct comparisons between studies or to conduct a quantitative meta-analysis. Additionally, the review did not assess the quality of individual studies, which is typical for scoping reviews but may limit the ability to evaluate the robustness of the reported findings. Finally, the focus on published literature may have introduced publication bias, potentially overlooking unpublished work or ongoing projects in the field of NLP application in healthcare.

Conclusions

In conclusion, whilst NLP techniques offer promising avenues for enhancing patient-centred care and quality improvement in healthcare, significant work remains to translate these technological advancements into meaningful clinical outcomes. Despite limitations, this review provides a comprehensive overview of the current state of NLP applications in analysing unstructured patient feedback across various healthcare settings. Whilst NLP techniques demonstrated potential in analysing large volumes of patient feedback efficiently, there was limited evidence of these insights translating into tangible clinical impacts or quality improvement initiatives. To realise the potential for NLP of UPF, future research must bridge the gap between academic interest and clinical impact. This calls for closer collaboration between NLP researchers and healthcare providers, study designs that demonstrate clear pathways to clinical application, and more effective methods for disseminating insights to healthcare professionals.

Acknowledgements

The study is funded by the European Commission's Horizon Europe Scheme (Reference number: 101057077) and involvement of the University of Manchester in this project is funded by the UKRI Innovate UK Horizon Guarantee scheme (Reference number: 10048830).

Author Contributions

Conceptualization: MB, LO, SL; Methodology: MB, LO, SL; Formal Analysis: AF, CL, MB, WT; Investigation: AF, CL, MB; Data Curation: AF, CL; Writing - Original Draft: AF, CL, MB, WT; Writing - Review & Editing: All authors; Visualization: AF; Supervision: MB, LO, SL, WT; Project Administration: MB; Funding Acquisition: MB, SL.

AI Use Statement

No generative artificial intelligence tools were used in the preparation or writing of this manuscript.

Data Availability

The datasets generated and analysed during this study, including the full data extraction spreadsheet, are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared

Abbreviations

PREMs: Patient Reported Experience Measures

UPF: Unstructured Patient feedback

NLP: Natural language processing

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

JB: Joanna Briggs Institute

CSRS: Cervical Spine Research Society

SVM: Support Vector Machines

CNN: Convolutional Neural Networks

RNN: Recurrent Neural Networks

LSTM: Long Short-Term Memory

LDA: Latent Dirichlet Allocation

NMF: Non-Negative Matrix Factorization

TCLDA: Topic Coherence-Based LDA

GSDMM: Gibbs Sampling Dirichlet Mixture Model

MTurk: Amazon Mechanical Turk



Multimedia Appendix 1: Search strategy

Full search strategy for the databases (MEDLINE, EMBASE, CINAHL, and Cochrane Database of Reviews). The search is built on the following concepts:

Public - lines 1 to 5

Feedback - lines 6 to 16

Healthcare setting - lines 17 to 22

NLP - line 23

1. Patients.mp. or exp Patients/
2. exp Consumer Behavior/
3. service adj3 user.mp.
4. client
5. public.mp.
6. exp Public Opinion/
7. exp Patient Satisfaction/
8. exp Patient-Centered Care/
9. exp Physician-Patient Relations/
10. exp Quality Assurance, Health Care/
11. exp Outcome Assessment, Health Care/
12. exp "Outcome and Process Assessment, Health Care"/ or exp Patient Outcome Assessment/
13. exp Outcome Assessment, Health Care/
14. exp Health Knowledge, Attitudes, Practice/ or exp "Surveys and Questionnaires"/ or exp Attitude to Health/
15. exp Patient Reported Outcome Measures/
16. exp Formative Feedback/ or exp Feedback/
17. exp Primary Health Care/ or exp "Delivery of Health Care"/
18. exp Dental Care/
19. exp Pharmacy Service, Hospital/ or exp Community Pharmacy Services/
20. exp Elective Surgical Procedures/
21. exp Hospitals/
22. exp Health Services/
23. natural language processing.mp. or exp Natural Language Processing/ or NLP.mp. or sentiment analysis.mp. or exp Sentiment Analysis/ or Latent Dirichlet allocation or Named entity recognition or topic modelling or text classification or keyword extraction
24. 1 or 2 or 3 or 4 or 5
25. 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16
26. 17 or 18 or 19 or 20 or 21 or 22
27. 23 and 24 and 25 and 26

Multimedia Appendix 2: Data extraction tool

Ref no	Title	Authors	Publication year	Journal	Data source(s) of unstructured feedback	Health care context	Number of texts used	Number of healthcare professionals rated	Type of health professional	Type of rating	NLP methods employed	Sentiment Analysis		Topic Modelling		Text Classification		Key findings
												Model used	Performance metrics	Model Used	Performance metrics	Model Used	Performance metrics	
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		

Multimedia Appendix 3: Summary tables

Summary table1: Sentiment analysis papers

Study	Study setting	Aim of study	Method used	Summary of findings
Yazdani et al (2023) [61]	- Collected 3600 feedback forms between March and October 2021 from the hospital website focusing on cancer patients.	- Detect cancer patients' sentiments on healthcare services in Persian language. - Combine sentiment analysis and topic modelling for healthcare service insights.	Sentistrength Topic modelling	- SA model achieved accuracies of 89.3%, 92.6%, and 90.8% in detecting sentiments. - Patients expressed dissatisfaction with appointment booking service. - System quality and interaction quality received majority of positive feedback.
Vasan et al. (2023) [33]	- Analysed 18546 reviews of 1240 otolaryngologists from the Healthgrades.com in March 2022.	- Analyse online reviews to uncover patterns among otolaryngologists' ratings. - Determine factors influencing sentiment and star scores in physician reviews.	VADER	- Younger and male otolaryngologists had higher sentiment and star scores. - Positive reviews associated with words like 'confident', 'kind', 'recommend', 'comfortable'. - Negative reviews had words like 'pain', 'rude', 'problems', 'wait'.
Tang et al. (2023) [34]	- Analysed 7638 reviews of 786 hand surgeons from healthgrades.com and webmd.com.	- Quantitatively analyse hand surgeons' reviews using sentiment analysis. - Report unbiased trends in words used to describe reviewed surgeons.	VADER	- Positive reviews linked to younger surgeons, good bedside manner. - No significant difference in review sentiment by provider sex.
Tang et al. (2022) [35]	- Collected 5121 reviews of 721 spine surgeons from healthgrades.com.	- Quantitatively analyse SRS surgeon reviews through sentiment analysis. - Determine relationship between demographic variables and average sentiment score.	VADER	- Sentiment scores correlate with star ratings, positive reviews for younger surgeons. - Positive reviews linked to compassionate care, efficient pain management. - Negative reviews associated with inability to relieve pain.
Tang et al. (2023) [36]	- Collected 8357 reviews of 480 spine surgeons from healthgrades.com.	- Analyse online reviews of spine surgeons for biases and trends. - Assess sentiment analysis impact on patient provider selection.	VADER	- Younger male surgeons rated higher. - Pain management crucial for positive reviews. - Positive reviews linked to personal characteristics.
Tang et al.	- Collected 2239 reviews of 177 spine surgeons from healthgrades.com.	- Quantitatively analyse CSRS surgeons' online reviews for biases	VADER	- Positive reviews linked to younger surgeons and behavioural attributes.

(2023) [37]		and trends. - Evaluate sentiment scores and word frequency in physician reviews.		- Pain management crucial for online reputation of CSRS surgeons. - Office environment impacts online reviews favourably.
Shah et al. (2021) [52]	- Collected 55612 reviews of 3430 doctors from ratemds, healthgrades.com, and vitals.	- Identify sentiment trends in physician rating websites. - Provide insights into patient opinions during the COVID-19 crisis.	Crystalfeel	- Negative emotions prevalent, fear, anger, and sadness during COVID-19. - Emerging topics focused on treatment process, awareness, and COVID-19 control. - Fading themes included panic buying, daily life impact, and bedside manner.
Shah et al. (2021) [53]	- Collected 152729 reviews of 10232 doctors from ratemds, healthgrades.com, and vitals.	- Investigate top concerns and sentiment dynamics in physician online reviews.	Crystalfeel SenticNET	- Sentiments shifted from positive to negative to positive during COVID-19. - ABSA captured sentiments towards healthcare quality aspects in online reviews.
Sewalk et al. (2018) [38]	- Collected 2759257 tweets related to healthcare experience between February 2013 and February 2017.	- Characterize patient experience sentiments on Twitter across the United States. - Examine differences in patient views of health care spatially and temporally.	VADER	- Metropolitan areas show higher positive tweets compared to non-metropolitan areas. - Decrease in tweet volume, increase in sentiment observed during study period.
Serrano-Guerrero et al. (2024) [21]	- Collected over 200,000 reviews from CareOpinion about healthcare professionals, covering a span of 5 years.	- Compare patient opinions on nurses and doctors for service improvement. - Assess patient needs and satisfaction to enhance healthcare services.	VADER	- Patients' opinions on nurses more positive than doctors.
Quinones et al. (2022) [39]	- Collected 6815 reviews of 1284 neurosurgeons in the USA from healthgrades.com in the period of July to August 2020.	- Investigate patient-physician interactions and key phrases in neurosurgeon reviews.	VADER	- No difference in sentiment scores based on sex of neurosurgeons. - Male neurosurgeons scored higher star ratings than females.
Park et al. (2023) [40]	- Collected 6400 reviews of 400 psychiatrists in the USA from healthgrades.com.	- Analyse patient reviews of psychiatrists on physician rating websites. - Identify factors influencing patient perception of psychiatrists. - Explore relationships between demographic factors, star ratings, and sentiment scores.	VADER	- Sentiment scores correlated with average star ratings of psychiatrists significantly. - Positive reviews mentioned "time" and "caring," negative reviews mentioned "medication." - Younger psychiatrists in the Northeast received higher star ratings.
Nawab et al.	- Collected 2830 reviews from Press Ganey	- Utilize NLP to mine free-text data for	Neural	- Doctors and nurses received most

(2020) [16]	Associates about a hospital in the USA.	valuable information.	networks	comments, room had negatives. - Negative comments on room, tests, and discharge categories. - Climate control, housekeeping, and noise levels major concerns in rooms.
Menendez et al. (2019) [57]	- Collected 254 reviews from Press Ganey Associates on an orthopaedic hospital in the USA.	- Analyse factors linked to negative comments in patient experiences post-surgery.	A 3 rd party tool (Press Ganey Associates' NLP tool)	- Patient-narrative analysis highlights care aspects critiqued by patients. - Negative comments themes include room condition, communication, and pain management. - Study population characteristics show age, sex, and insurance status distribution.
Li et al. (2022) [54]	- Collected 58600 reviews about doctors in China from haodf.com	- Investigate changes in patient attitudes towards doctors during COVID-19 pandemic. - Analyse factors influencing improvements in doctor-patient relationships post-pandemic.	A 3 rd party tool (Baidu Cloud)	- Improved doctor-patient relationships post-COVID-19 in China. - Decrease in negative reviews, shift towards positive attitudes post-pandemic. - Sentiment analysis showed slight increase in positive comments during the pandemic.
Levy et al. (2023) [41]	- Collected 6606 reviews from healthgrades.com about 1854 urologists in the USA.	- Identify factors affecting online reviews for urologists nationwide.	VADER	- Older age doctors linked to more negative reviews in urologists. - Positive reviews tied to physician's attitude and character phrases. - Negative reviews associated with long wait times and patient pain.
Langerhuizen et al. (2021) [56]	- Collected 11614 reviews of orthopaedic surgeons from Yelp in the USA.	- Identify themes in online reviews of orthopaedic surgeons. - Analyse patient experience based on review ratings and tones.	3 rd party (IBM Watson)	- Themes varied by rating category, including logistics, care, trust, and recommendation. - Positive reviews were shorter, with joy and confidence tones. - Negative reviews had sadness and tentative tones.
Khanbhai et al. (2022) [43]	- A total of 69285 reviews were collected at a large London NHS Trust hospital.	- Extract themes and sentiment from patient experience comments using NLP. - Identify issues in transitions of care across different healthcare settings.	SVM	- Negative sentiments include 'discharge', 'appointment', and 'home'. - Tri-grams highlight issues like 'seeing different doctor' and 'information aftercare lacking'. - NLP identifies problems with care transitions

				and continuity effectively.
Khaleghparast et al. (2023) [49]	- A total of 822 reviews were collected from a cardiovascular hospital in Iran.	- Perform sentiment analysis on patients' messages using lexicon-based and ML methods. - Identify positive and negative comments and determine hospital ward and staff names.	- AdaBoost - Decision Tree - Logistic Regression - Multilayer Perceptron - Naïve Bayes	- Best classifier: Multinomial Naïve Bayes with MF + TFIDF feature vector. - Previous studies used SA and ML to classify patients' comments effectively
Kao et al. (2015) [50]	- Collected 63 reviews of an academic multidisciplinary pain centre.	- Develop machine-mediated classification of patient feedback using NLP. - Optimize patient experience monitoring with low cost and resource burden.	Recurrent Neural Networks (RNN)	- The optimal cut-off for classifying the feedback as Negative is a threshold of 68.5% or more of the sentences being Negative. - There is moderate agreement between machine-mediated rating with single-rater coding, with Kappa of 0.57 and AUC of 0.79
Jung et al. (2015) [76]	- Collected 173748 reviews of a paediatric hospital in South Korea from online communities	- Identify hospital service quality factors and trends from online health communities.	Keyword-based approach	- Service quality factor detection and hospital name extraction achieved average F1 scores of 91% and 78%, respectively
Jiménez-Zafra et al. (2019) [44]	- A total of 743 reviews of doctors with 34 specialties were collected from Masquemedicos, a medical community in Spain. They also collected reviews on drugs.	- Examine sentiment expression in medical forums for sentiment analysis methods. - Analyse language to determine best approach for sentiment analysis.	- SVM - iSOL Lexicon	- Drug reviews challenging, informal language with specific terminology. - Physicians' reviews informal, easier to classify than drug reviews. - Reviews about drugs longer, opinions about physicians have more adjectives.
Huppertz and Otto (2018) [45]	- Collected 57985 reviews from 113 hospitals' Facebook page in the USA.	- Social media feedback predicts hospital HCAHPS scores accurately. - Sentiment analysis on Facebook comments enhances patient experience assessment.	- Naïve Bayes - KNN - SVM	- Both number of stars and the number of positive comments posted on hospitals Facebook Reviews sections were associated with higher overall ratings and willingness to recommend the hospital. - The findings suggest that patients' informal comments help predict a hospital's formal measures of patient experience.
Hu et al. (2019) [55]	- Collected approximately 29 million records from WeChat and Qzone in China about healthcare services.	- Identify content volume and sentiment polarity of healthcare services in China. - Analyse public perceptions on healthcare services using social media	3 rd party (Tencent NLP platform)	- Patient safety most discussed topic, negative sentiments in doctor-patient relationship. - Sentiment analyses showed 36.1%, 16.4%, and 47.4% of positive, neutral, and negative

		data.		emotions, respectively. - Neutral disposition was found to be the highest (30.4%) in the contents on appointment-booking services.
Hawkins et al. (2016) [48]	- Collected 404065 tweets about hospitals in the USA	- Assess Twitter as a measure of patient-perceived quality in hospitals. - Compare patient sentiments on Twitter with established quality measures.	- Naïve Bayes - SVM	- Positive sentiment on Twitter, hospitals interact with users for feedback. - Hospitals with more tweets had higher sentiment, no link to readmission.
Gui and He (2021) [51]	- Collected 201246 reviews on Yelp about healthcare services	- Propose joint learning framework for aspect extraction and sentiment classification	- Convolution Neural Network (CNN) - Recurrent Neural Network (RNN) - Long-Short Terms Memory (LSTM)	- Aspect-based model outperforms existing neural models for sentiment classification. - Aspect-based model extracts coherent aspect topics and learns good document representations. - Aspect-based model visualisation shows word and sentence-level attention weights in aspect-based.
Gour and Kumari (2021) [58]	- Collected 9492 reviews on hospitals in India from Various websites such as Facebook, Quora, MouthShut, Google reviews and Just Dial.	- Develop fuzzy sentiment analysis model for hospital patient reviews. - Analyse reviews to help hospitals find pros and cons.	Naïve Bayes	- Traditional vs. fuzzy sentiment analysis showed significant differences in review sentiments. - Trust and positive emotions were prominent in the hospital reviews. - Negative emotions were more prevalent in fuzzy sentiment analysis results.
Pandey et al. (2023) [42]	- Collected 178 reviews of 34 General Practitioners (GPs) in Northamptonshire, UK from NHS website.	- Improve patient experience in healthcare. - Classify GPs based on feedback. - Enhance patient feedback processes.	VADER	- Classifier grouped GPs into gold, silver, and bronze categories. - 10-fold cross-validation test mode for data set classification.
Almorox et al. (2022) [80]	- Collected 2195 reviews of 242 care homes in the UK from carehome.co.uk.	- Examine sentiment changes in care home reviews during COVID-19. - Assess impact of integrated care on residents' and carers' experiences.	AFINN	- Care home sentiment increased during COVID-19, especially with integration terms. - Staff role crucial in driving sentiment for residents and relatives.
Alemi et al. (2012) [46]	- Collected 995 reviews of 200 Paediatricians and obstetricians/gynaecologists in the USA from ratemds.com	- To demonstrate sentiment analysis feasibility in real-time satisfaction surveys.	- Decision trees - bagging with decision trees	- Sentiment analysis classifies comments into complaints or praise accurately. - Majority of patient reviews and comments

		- To classify patients' reasons for dissatisfaction using sentiment analysis.	- Naïve Bayes - SVM	are positive and constructive.
Agarwal et al. (2022) [47]	- Collected approximately 10000 tweets in India about healthcare services.	- Extracting real-time Twitter data concerning three healthcare sub-domains: COVID-19 vaccine, post-COVID-19 health factors, and healthcare service providers. - Using machine learning techniques to classify and analyse the sentiments and narratives expressed by social media users regarding these sub-domains.	- Support Vector Machine (SVM) - Logistic Regression - Random Forest - Multinomial Naive Bayes (MNB) -Long Short-Term Memory (LSTM)	- Sentiment analysis of the collected tweets provided insights into public opinion regarding the effectiveness of COVID-19 vaccines, long-term health impacts post-COVID, and the performance of healthcare service providers. - SVM model showed highest accuracy of 82.67% compared to others.

Summary table2: Topic modelling papers

Study	Study setting	Aim of study	Method used	Summary of findings
Yazdani et al (2023) [61]	- Collected 3600 feedback forms between March and October 2021 from the hospital website focusing on cancer patients.	- Detect cancer patients' sentiments on healthcare services in Persian language. - Combine sentiment analysis and topic modelling for healthcare service insights.	LDA	- The topic "Metastasis" exhibited lower sentiment scores compared to other topics. - Cancer patients expressed dissatisfaction with the current appointment booking service. - Topics such as "Good experience," "Affable staff", and "Chemotherapy" garnered higher sentiment scores.
Stokes et al. (2021) [62]	- Collected 8133 Yelp reviews of 1383 mental health treatment facilities.	- Evaluate correlation between review themes, facility characteristics, and ratings.	LDA	- Positive themes: caring staff, non-pharmacologic treatment. - Negative themes: rude staff, safety and abuse. - Patient satisfaction linked to staff-patient communication. - Reviews highlight concerns on safety, communication, and staff attitudes.
Ranard et al.	- Collected 16862	- Compare Yelp	LDA via MALLET	- Yelp ratings correlate with HCAHPS overall ratings.

(2016) [7]	Yelp reviews of 1352 hospitals.	reviews with Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey domains. - Identify topics correlating with Yelp ratings and HCAHPS overall ratings.		- Yelp reviews cover more topics than HCAHPS survey. - Yelp topics reveal major drivers of patient experiences.
Lin et al. (2020) [27]	- Collected online reviews of 204,751 dentists from HealthGrades.com	- Characterize online reviews of dental patients in the United States. - Examine association between dentist characteristics and patient ratings.	A semi-automated NLP method to perform text mining (bigram and trigram), focusing on identifying relevant concepts. Then manually assessing the collocations manually by 2 raters.	- Higher ratings were associated with female dentists. - Patient ratings linked to dentist characteristics and patient experience topics. - Topics that corresponded to CAHPS measures, including discomfort (painful/painless root canal or deep cleaning), and ethics (high-pressure sales, and unnecessary dental work)
Lester and Chui (2017) [65]	- A total of 1,421 user reviews from 423 unique pharmacies were collected from Yelp.	- Identify patient topics at community pharmacies and their impact on Yelp ratings. - Improve patient experience and provide actionable feedback for pharmacies.	LDA	- Topics from Yelp reviews inform pharmacy improvement for patient experience. - Identified four discernible topics including long wait times for prescriptions, helpful pharmacy staff, good store environment, and issues filling medications.
Graves et al. (2018) [81]	- Collected 42752 Yelp reviews on hospitals in the US regarding the pain management and opioids.	- Characterize Yelp reviews on pain management and opioid experiences. - Identify themes impacting satisfaction and patient experiences with opioids.	Differential language analysis (DLA)	- Patients describe negative factors like long wait times and poor pain control. - Positive aspects include attentive medical staff. - Reviews show perceptions of excessive opioid use and inappropriate care.

Chekijian et al. (2021) [82]	- Collected 6406 anonymised reviews from Press Ganey about hospitals.	- Understand patient experience during COVID-19 using sentiment analysis and topic modelling. - Identify key trends in sentiment and topics related to patient experience.	LDA	- Patient concerns shifted pre, during, and post-COVID-19 with varying sentiments. - Topics included safety, family involvement, and system issues during COVID-19. - Positive sentiments decreased during COVID-19 but increased post-COVID-19.
Chan et al. (2022) [83]	- Collected 619 Yelp reviews on crisis pregnancy centres in California.	- Analyse patient experiences and motivations at California Crisis Pregnancy Centres (CPCs). - Evaluate crowdsourced reviews of CPCs using Yelp.	Thematic coding validated by meaning extraction method.	- Motivations for CPC care: pregnancy confirmation, emotional support, abortion care. - CPC experiences: mixed, faith-based practices, inaccurate medical information, lack of transparency. - Patients seek CPCs due to gaps in healthcare coverage, free materials. - CPCs serve socially vulnerable groups, including uninsured and low-income patients.
Cammel et al. (2020) [74]	- Collected 38664 patients' experience data from 2 hospitals.	- Categorize patient responses using unsupervised topic modelling. - Create combined sentiment and frequency measure for improvement prioritization. - Assess model transferability on free-text responses from another hospital.	- Non-negative matrix factorization (NMF) - Latent Dirichlet Allocation (LDA)	- the first hospital resulted in 127 topics and 294 n-grams. - The indicator 'impact' revealed n-grams to celebrate (15.3%), improve (8.8%), and monitor (16.7%). - Between-hospitals, most topics (69.7%) were similar, but 32.2% of topics for hospital 1 and 29.0% of topics for hospital 2 were unique.
Agarwal et al. (2019) [59]	- Collected 100949 Yelp reviews on urgent care centres.	- Compare online reviews of EDs and urgent care centres. - Identify topics correlated with 1-and	LDA	- Patients value comfort, professionalism, facilities, and staff interactions. - Themes associated with 5-star reviews among EDs and urgent care centres were similar for comfort, professionalism, facilities, paediatric care, and staff interactions.

		5-star ratings in reviews.		- Themes associated with 1-star reviews among EDs and urgent care centres were similar for communication, telephone experience, waiting, billing, pain management, and diagnostic testing. - Themes unique to 5-star ED reviews included bedside manner, care for family members, and access. Themes unique to 5-star urgent care centre reviews were based on recommendation and prescription refills. Themes unique to 1-star ED reviews were service and speed of care. Themes unique to 1-star urgent care centre reviews were lack of confidence and reception experience.
--	--	----------------------------	--	---

Summary table 3: Text classification papers

Study	Study setting	Aim of study	Method used	Summary of findings
Rajagopalan et al. (2022) [84]	- Collected 1390 reviews on various specialties of doctors in the USA	- Develop quantitative patient experience measures using natural language processing. - Address ceiling effects in patient-reported experience measures.	- Linguistic Inquiry and Word Count (LIWC) - Thematic analysis	- Analysis identified 36 emotional and linguistic constructs associated with ordinal rating of likelihood to recommend - LIWC used to summarize text comments for likelihood to recommend ratings. - the score developed from thematic analysis both had a relatively normal distribution and limited or no ceiling effect.
Parikh et al. (2024) [73]	- Collected 2292 reviews about radiology of a hospital department in the USA.	- Describe effects of demographics and examination factors on MRI experience. - Understand patient feedback for patient-centred changes.	- Sentence BERT - Multilabel XGBoost	- Male gender linked to positive ratings, shorter exams, and punctuality. - Longer wait times and exams associated with negative patient ratings. - Positive feedback themes: excellent service, on-time appointments, comfortable IV placement. - Negative feedback themes: long wait times, poor communication, physical discomfort.
He et al. (2020) [72]	- Collected 1,065,631 reviews on 102,540 doctors in the USA from Vitals website.	- Identify patient concerns in online physician reviews through qualitative analysis. - Improve healthcare providers' ratings	- Manually labelled 600 reviews and then trained a classifier.	- Relationship is a key concern in patient reviews. - Management issues trigger negative reviews more than clinical performance. - Fine-grained aspects include timing, access, finance, and communication.

		based on consumer feedback.		
--	--	-----------------------------	--	--

References

1. Burt J, Campbell J, Abel G, Aboulghate A, Ahmed F, Asprey A. Improving patient experience in primary care: a multimethod programme of research on the measurement and improvement of patient experience. Programme Grants for Applied Research. 2017;5(9). doi: <https://doi.org/10.3310/pgfar05090>.
2. Fenton JJ, Jerant AF, Bertakis KD, Franks P. The cost of satisfaction: a national study of patient satisfaction, health care utilization, expenditures, and mortality. Archives of Internal Medicine. 2012;172(5):405-11. doi: <https://doi.org/10.1001/archinternmed.2011.1662>.
3. Kessler DP, Mylod D. Does patient satisfaction affect patient loyalty? International Journal of Health Care Quality Assurance. 2011;24(4):266-73. doi: <https://doi.org/10.1108/09526861111125570>.
4. Gleeson H, Calderon A, Swami V, Deighton J, Wolpert M, Edbrooke-Childs J. Systematic review of approaches to using patient experience data for quality improvement in healthcare settings. BMJ open. 2016;6(8):e011907. doi: <https://doi.org/10.1136/bmjopen-2016-011907>.
5. Lu Z, Sim J-A, Wang JX, Forrest CB, Krull KR, Srivastava D, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. Journal of Medical Internet Research. 2021;23(11):e26777. doi: <https://doi.org/10.2196/26777>.
6. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Quality & Safety. 2013;22(3):251-5. doi: <https://doi.org/10.1136/bmjqs-2012-001527>.
7. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, et al. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. Health Affairs. 2016;35(4):697-705. doi: <https://doi.org/10.1377/hlthaff.2015.1030>.
8. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. BMJ Quality & Safety. 2016;25(8):604-14. doi: <https://doi.org/10.1136/bmjqs-2015-004063>.
9. Reimann S, Strech D. The representation of patient experience and satisfaction in physician rating sites. A criteria-based analysis of English- and German-language sites. BMC Health Services Research. 2010;10(Dec):332. doi: <https://doi.org/10.1186/1472-6963-10-332>.
10. Boylan AM, Williams V, Powell J. Online patient feedback: a scoping review and stakeholder consultation to guide health policy. Journal of Health Services Research and Policy. 2020;25(2):122-9. doi: <https://doi.org/10.1177/1355819619870837>.
11. Kim HL, Houman J, Daskivich TJ, Fuller G, Spiegel B, Black JT. Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. Journal of the American Medical Informatics Association. 2017;25(4):401-7. doi: <https://doi.org/10.1093/jamia/ocx083>.
12. Chen J, Presson A, Zhang C, Ray D, Finlayson S, Glasgow R. Online physician review websites poorly correlate to a validated metric of patient satisfaction. Journal of Surgical Research. 2018;227:1-6. doi: <https://doi.org/10.1016/j.jss.2018.01.037>.
13. Griffiths A, Leaver MP. Wisdom of patients: predicting the quality of care using aggregated patient feedback. BMJ Quality & Safety. 2018;27(2):110-8. doi: <https://doi.org/10.1136/bmjqs-2017-006847>.

14. Trehan SK, Daluiski A. Online Patient Ratings: Why They Matter and What They Mean. *The Journal of Hand Surgery*. 2016;41(2):316-9. doi: <https://doi.org/10.1016/j.jhsa.2015.04.018>.
15. Liddy ED. Natural language processing. Syracuse University, 2001.
16. Nawab K, Ramsey G, Schreiber R. Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback. *Applied Clinical Informatics*. 2020;11(2):242-52. doi: <https://doi.org/10.1055/s-0040-1708049>.
17. Guda N. Analyzing the Extent to which Gender Bias Exists in News Articles Using Natural Language Processing Techniques. *Journal of Student Research*. 2023;12(1):1-12. doi: 10.47611/jsrhs.v12i1.
18. Belz A. A Metrological Perspective on Reproducibility in NLP*. *Computational Linguistics*. 2022;48(4):1125-35. doi: 10.1162/coli_a_00448.
19. Alexander G, Bahja M, Butt GF. Automating Large-scale Health Care Service Feedback Analysis: Sentiment Analysis and Topic Modeling Study. *JMIR Med Inform*. 2022 Apr 11;10(4):e29385. PMID: 35404254. doi: 10.2196/29385.
20. Farrell MJ, Brierley L, Willoughby A, Yates A, Mideo N. Past and future uses of text mining in ecology and evolution. *Proceedings of Biological Sciences*. 2022 May 25;289(1975):20212721. PMID: 35582795. doi: 10.1098/rspb.2021.2721.
21. Jesus Serrano-Guerrero MB-D, Francisco Chiclana,, Olivas FPRJA. How satisfied are patients with nursing care and why? A comprehensive study based on social media and opinion mining. *Informatics for Health and Social Care*. 2024;311(Aug):18-38. doi: <https://doi.org/10.1080/17538157.2023.2297307>.
22. Scharkow M. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*. 2013;47(2):761-73. doi: <https://doi.org/10.1007/s11135-011-9545-7>.
23. Canini K, Shi L, Griffiths T, editors. Online inference of topics with latent Dirichlet allocation. *Artificial Intelligence and Statistics*; 2009: PMLR.
24. van Buchem MM, Neve OM, Kant IM, Steyerberg EW, Boosman H, Hensen EF. Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM). *BMC Medical Informatics and Decision Making*. 2022;22(1):1-11. doi: <https://doi.org/10.1186/s12911-022-01923-5>.
25. Rahim AIA, Ibrahim MI, Musa KI, Chua SL. Facebook reviews as a supplemental tool for hospital patient satisfaction and its relationship with hospital accreditation in malaysia. *International Journal of Environmental Research and Public Health*. 2021;18(14):20-39. doi: <https://doi.org/10.3390/ijerph18147454>.
26. Gray BM, Vandergrift JL, Gao G, McCullough JS, Lipner RS. Website Ratings of Physicians and Their Quality of Care. *JAMA Internal Medicine*. 2015;175(2):291-3. doi: <https://doi.org/10.1001/jamainternmed.2014.6291>.
27. Lin Y, Hong YA, Henson BS, Stevenson RD, Hong S, Lyu T, et al. Assessing patient experience and healthcare quality of dental care using patient online reviews in the United States: mixed methods study. *Journal of Medical Internet Research*. 2020;22(7):e18652. doi: <https://doi.org/10.2196/18652>.
28. Ponathil A, Khasawneh A, Byrne K, Chalil Madathil K. Factors affecting the choice of a dental care provider by older adults based on online consumer reviews. *IISE Transactions on Healthcare Systems Engineering*. 2020;11(1):51-69. doi:

<https://doi.org/10.1080/24725579.2020.1854394>.

29. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health & Care Informatics*. 2021;28(1):e100262. doi: <https://doi.org/10.1136/bmjhci-2020-100262>.
30. Peters MDJ, Godfrey C, Mclnerney P, Munn Z, Tricco AC, Khalil H. JBI Manual for Evidence Synthesis. In: Aromataris E, Z M, editors.: JBI; 2020.
31. Rayyan. Rayyan – Intelligent Systematic Review - Rayyan. 2024 [cited 2024 1st November]; Available from: <https://www.rayyan.ai/>.
32. Hutto C, Gilbert E, editors. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*; 2014.
33. Vasan V, Cheng CP, Lerner DK, Vujovic D, van Gerwen M, Iloreta AM. A natural language processing approach to uncover patterns among online ratings of otolaryngologists. *The Journal of laryngology and Otology*. 2023;137(12):1384-8. doi: <https://dx.doi.org/10.1017/S0022215123000476>.
34. Tang JE, Arvind V, White CA, Dominy C, Kim JS, Cho SK, et al. Using Sentiment Analysis to Understand What Patients Are Saying About Hand Surgeons Online. *Hand*. 2023;18(5):854-60. doi: <https://dx.doi.org/10.1177/15589447211060439>.
35. Tang JE, Arvind V, White CA, Dominy C, Kim JS, Cho SK. What are patients saying about you online? A sentiment analysis of online written reviews on Scoliosis Research Society surgeons. *Spine Deformity*. 2022;10(2):301-6. doi: <https://dx.doi.org/10.1007/s43390-021-00419-y>.
36. Tang JE, Arvind V, Dominy C, White CA, Cho SK, Kim JS. How Are Patients Reviewing Spine Surgeons Online? A Sentiment Analysis of Physician Review Website Written Comments. *Global Spine Journal*. 2023;13(8):2107-14. doi: <https://dx.doi.org/10.1177/21925682211069933>.
37. Tang J, Arvind V, White CA, Dominy C, Cho S, Kim JS. How are Patients Describing You Online? A Natural Language Processing Driven Sentiment Analysis of Online Reviews on CSRS Surgeons. *Clinical Spine Surgery*. 2023;36(2):E107-E13. doi: <https://dx.doi.org/10.1097/BSD.0000000000001372>.
38. Sewalk KC, Tuli G, Hsuen Y, Brownstein JS, Hawkins JB. Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study. *Journal of Medical Internet Research*. 2018;20(10):84-. doi: <https://doi.org/10.2196/10043>.
39. Quinones A, Tang J, Vasan V, Li T, Li A, Durbin J, et al. Trends in Online Patient Perspectives of Neurosurgeons: A Sentiment Analysis. *Journal of Neurosurgery*. 2022;136(5):45-62. doi: <https://dx.doi.org/10.3171/2022.5.JNS.AANS2022abstracts>.
40. Park SH, Cheng CP, Buehler NJ, Sanford T, Torrey W. A sentiment analysis on online psychiatrist reviews to identify clinical attributes of psychiatrists that shape the therapeutic alliance. *Frontiers in Psychiatry*. 2023 2023;14:1174154. doi: <https://dx.doi.org/10.3389/fpsy.2023.1174154>.
41. Levy M, Tang JE, Chin CP, Ravivarapu KT, Gologorsky R, White CA, et al. Sentiment Analysis of Online Written Reviews for a National Urology Cohort Illustrates Important Factors for Patient Satisfaction. *Journal of Urology*. 2023;209(Supplement 4):e58. doi: <https://dx.doi.org/10.1097/JU.0000000000003217.13>.
42. Pandey AR, Seify M, Okonta U, Hosseinian-Far A. Advanced Sentiment Analysis for Managing and Improving Patient Experience: Application for General Practitioner (GP)

Classification in Northamptonshire. *International Journal of Environmental Research and Public Health*. 2023;20(12):6119. doi: <https://doi.org/10.3390/ijerph20126119>.

43. Khanbhai M, Warren L, Symons J, Flott K, Harrison-White S, Manton D, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. *International Journal of Medical Informatics*. 2022;157(Jan):104642. doi: <https://dx.doi.org/10.1016/j.ijmedinf.2021.104642>.

44. Jimenez-Zafra SM, Martin-Valdivia MT, Molina-Gonzalez MD, Urena-Lopez LA. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. *Artificial Intelligence in Medicine*. 2019;93(Jan):50-7. doi: <https://dx.doi.org/10.1016/j.artmed.2018.03.007>.

45. Huppertz JW, Otto P. Predicting HCAHPS scores from hospitals' social media pages: A sentiment analysis. *Health Care Management Review*. 2018;43(4):359-67. doi: <https://dx.doi.org/10.1097/HMR.0000000000000154>.

46. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Quality management in health care*. 2012;21(1):9-19. doi: <https://dx.doi.org/10.1097/QMH.0b013e3182417fc4>.

47. Agrawal S, Jain SK, Sharma S, Khatri A. COVID-19 Public Opinion: A Twitter Healthcare Data Processing Using Machine Learning Methodologies. *International journal of environmental research and public health*. 2022;20(1):630-43. doi: <https://dx.doi.org/10.3390/ijerph20010432>.

48. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Quality & Safety*. 2016;25(6):404-13. doi: <https://dx.doi.org/10.1136/bmjqs-2015-004309>.

49. Khaleghparast S, Maleki M, Hajianfar G, Soumari E, Oveisi M, Gol, et al. Development of a patients' satisfaction analysis system using machine learning and lexicon-based methods. *BMC Health Services Research*. 2023;23(1):280. doi: <https://dx.doi.org/10.1186/s12913-023-09260-7>.

50. Kao M, Leong M, Prasad R, Cullen A, Pacht T, Olson G, et al. Stanford Patient Experience Questionnaire (SPEQ): Machine-mediated classification of patient experience feedback using natural language processing. *Journal of Pain*. 2015;16(4):S37. doi: <https://doi.org/10.1016/j.jpain.2015.01.160>.

51. Gui L, He Y. Understanding patient reviews with minimum supervision. *Artificial Intelligence in Medicine*. 2021;120(October):102160. doi: <https://doi.org/10.1016/j.artmed.2021.102160>.

52. Shah AM, Yan X, Qayyum A, Naqvi RA, Shah SJ. Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach. *International Journal of Medical Informatics*. 2021;149(May):104434. doi: <https://dx.doi.org/10.1016/j.ijmedinf.2021.104434>.

53. Shah AM, Naqvi RA, Jeong OR. Detecting topic and sentiment trends in physician rating websites: Analysis of online reviews using 3-wave datasets. *International Journal of Environmental Research and Public Health*. 2021;18(9):4743. doi: <https://dx.doi.org/10.3390/ijerph18094743>.

54. Li J, Pang PCI, Xiao Y, Wong D. Changes in Doctor-Patient Relationships in China during COVID-19: A Text Mining Analysis. *International Journal of Environmental Research and Public Health*. 2022;19(20):13446. doi: <https://dx.doi.org/10.3390/ijerph192013446>.

55. Hu G, Han X, Zhou H, Liu Y. Public Perception on Healthcare Services: Evidence from

- Social Media Platforms in China. *International Journal of Environmental Research and Public Health*. 2019;16(7):1273. doi: <https://dx.doi.org/10.3390/ijerph16071273>.
56. Langerhuizen DWG, Brown LE, Doornberg JN, Ring D, Kerkhoffs GMMJ, Janssen SJ. Analysis of Online Reviews of Orthopaedic Surgeons and Orthopaedic Practices Using Natural Language Processing. *The Journal of the American Academy of Orthopaedic Surgeons*. 2021;29(8):337-44. doi: <https://dx.doi.org/10.5435/JAAOS-D-20-00288>.
57. Menendez ME, Shaker J, Lawler SM, Ring D, Jawa A. Negative Patient-Experience Comments After Total Shoulder Arthroplasty. *The Journal of bone and joint surgery*. 2019;101(4):330-7. doi: <https://dx.doi.org/10.2106/JBJS.18.00695>.
58. Gour A, Kumari S. A 360-Degree View of a Hospital by Analysing Patient's Online Reviews Using Fuzzy Sentiment Analysis. *Journal of Health Management*. 2021;23(3):549-57. doi: <https://doi.org/10.1177/09720634211032017>.
59. Agarwal AK, Mahoney K, Lanza AL, Klinger EV, Asch DA, Fausti N, et al. Online Ratings of the Patient Experience: Emergency Departments Versus Urgent Care Centers. *Annals of emergency medicine*. 2019;73(6):631-8. doi: <https://dx.doi.org/10.1016/j.annemergmed.2018.09.029>.
60. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of machine Learning research*. 2003;3(Jan):993-1022.
61. Yazdani A, Shamloo M, Khaki M, Nahvijou A. Use of sentiment analysis for capturing hospitalized cancer patients' experience from free-text comments in the Persian language. *BMC Medical Informatics & Decision Making*. 2023;23(1):1-14. doi: <https://doi.org/10.1186/s12911-023-02358-2>.
62. Stokes DC, Kishton R, McCalpin HJ, Pelullo AP, Meisel ZF, Beidas RS, et al. Online Reviews of Mental Health Treatment Facilities: Narrative Themes Associated With Positive and Negative Ratings. *Psychiatric Services*. 2021;72(7):776-83. doi: <https://dx.doi.org/10.1176/appi.ps.202000267>.
63. Quality AfHRA. CAHPS Measures of Patient Experience. 2024 [cited 2024 1st November]; Available from: <https://www.ahrq.gov/cahps/consumer-reporting/measures/index.html>.
64. Services CfMM. HCAHPS: Patients' Perspectives of Care Survey. 2024 [cited 2024 1st November]; Available from: <https://www.cms.gov/medicare/quality/initiatives/hospital-quality-initiative/hcahps-patients-perspectives-care-survey>.
65. Lester C, Chui M. Evaluating the patient experience at community pharmacies using yelp reviews. *Journal of the American Pharmacists Association*. 2017;57(3):30-45. doi: <https://dx.doi.org/10.1016/j.japh.2017.04.011>.
66. Blei DM, Lafferty JD, editors. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*; 2006.
67. Yin J, Wang J, editors. A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2014.
68. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401(6755):788-91. doi: <https://doi.org/10.1038/44565>.
69. Chuang J, Manning CD, Heer J, editors. Termite: Visualization techniques for assessing textual topic models. *Proceedings of the international working conference on advanced visual interfaces*; 2012.
70. Li Q, Peng H, Li J, Xia C, Yang R, Sun L, et al. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

2022;13(2):1-41. doi: <https://doi.org/10.1145/3495162>.

71. KNIME. KNIME Analytics Platform. [cited 2024 1st June]; Available from: <https://www.knime.com/knime-analytics-platform>.

72. He L, He C, Wang Y, Hu Z, Zheng K, Chen Y. What Do Patients Care About? Mining Fine-grained Patient Concerns from Online Physician Reviews Through Computer-Assisted Multi-level Qualitative Analysis. AMIA Annual Symposium proceedings AMIA Symposium. 2020;2020:544-53.

73. Parikh P, Klanderman M, Teck A, Kunzelman J, Banerjee I, DeYoung D, et al. Effects of patient demographics and examination factors on patient experience in outpatient MRI appointments. Journal of the American College of Radiology. 2024;21(4):601-8. doi: <https://doi.org/10.1016/j.jacr.2023.02.032>.

74. Cammel SA, De Vos MS, van Soest D, Hettne KM, Boer F, Steyerberg EW, et al. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. BMC Medical Informatics and Decision Making. 2020;20(1):97. doi: <https://doi.org/10.1186/s12911-020-1104-5>.

75. Alemi F, Jasper H. An alternative to satisfaction surveys: let the patients talk. Quality management in health care. 2014;23(1):10-9. doi: <https://dx.doi.org/10.1097/QMH.0000000000000014>.

76. Jung Y, Hur C, Jung D, Kim M. Identifying key hospital service quality factors in online health communities. Journal of Medical Internet Research. 2015;17(4):e90. doi: <https://dx.doi.org/10.2196/jmir.3646>.

77. Koroteev MV. BERT: a review of applications in natural language processing and understanding. arXiv preprint. 2021. doi: <https://doi.org/10.48550/arXiv.2103.11943>.

78. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv preprint. 2023;1(2).

79. Brown B, Gude WT, Blakeman T, van der Veer SN, Ivers N, Francis JJ, et al. Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. Implementation science : IS. 2019 2019/04/26;14(1):40. doi: 10.1186/s13012-019-0883-5.

80. Almorox EG, Stokes J, Morciano M. Has COVID-19 changed carer's views of health and care integration in care homes? A sentiment difference-in-difference analysis of on-line service reviews. Health Policy. 2022;126(11):1117-23. doi: [doi:10.1016/j.healthpol.2022.08.010](https://doi.org/10.1016/j.healthpol.2022.08.010).

81. Graves RL, Goldshear J, Perrone J, Ungar L, Klinger E, Meisel ZF, et al. Patient narratives in Yelp reviews offer insight into opioid experiences and the challenges of pain management. Pain Management. 2018;8(2):95-104. doi: <https://doi.org/10.2217/pmt-2017-0050>.

82. Chekijian S, Li H, Fodeh S. Emergency care and the patient experience: Using sentiment analysis and topic modeling to understand the impact of the COVID-19 pandemic. Health and Technology. 2021;11(5):1073-82. doi: <https://doi.org/10.1007/s12553-021-00585-z>.

83. Chan E, Korotkaya Y, Osadchiy V, Sridhar A. Patient Experiences at California Crisis Pregnancy Centers: A Mixed-Methods Analysis of Online Crowd-Sourced Reviews, 2010-2019. Southern Medical Journal. 2022;115(2):144-51. doi: <https://doi.org/10.14423/smj.0000000000001353>.

84. Rajagopalan D, Thomas J, Ring D, Fatehi A. Quantitative Patient-Reported Experience

Measures Derived From Natural Language Processing Have a Normal Distribution and No Ceiling Effect. Qual Manag Health Care. 2022;31(4):210-8. doi: <https://doi.org/10.1097/qmh.0000000000000355>.

Preprint
JMIR Publications

Supplementary Files

Untitled.

URL: <http://asset.jmir.pub/assets/ea03b59b17731624fd63c355aac73fe4.docx>

Multimedia Appendixes

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
URL: <http://asset.jmir.pub/assets/11633f1fa433553c24707b9f14f2bc7f.pdf>

CONSORT (or other) checklists

Untitled.

URL: <http://asset.jmir.pub/assets/712b371ae99caf0459bc9d2d51e58bf5.pdf>