

Natural Language Processing of Sentiments Identified in Patient Comments Associated with Less Than Top-Rated Care

Journal of Patient Experience
Volume 12: 1-5
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23743735251323677
journals.sagepub.com/home/jpx



Ali Azarpey, MD¹ , Jacob Thomas, PhD¹, David Ring, MD, PhD¹ ,
and Orrin Franko, MD²

Abstract

Background: Natural language processing (NLP) analysis of patient comments about their care can inform improvement initiatives. **Objective:** We used NLP to quantify sentiments and identify topics in patient comments associated with submaximal ratings of experience. **Methods:** Using a set of 1117 patient comments associated with ratings 1-4 out of 5 from a commercial source, we analyzed associated sentiments measured by Linguistic Inquiry and Word Count software and associated themes using topic modeling. **Results:** In the sentiment analysis, positive sentiments were associated with better numerical ratings while word count, numbers, ethnicity, and negative tones were associated with lower ratings. Topics of “listening, concern, and collaboration” were associated with 1-star ratings and “logistics” and “pain” with 4-star ratings. **Conclusion:** The finding that NLP analysis of comments from submaximal patient ratings of experience is consistent with evidence that the worst ratings are associated with relationship issues and more moderate ratings are associated with process issues affirms the ability of NLP to analyze large amounts of patient comments to identify opportunities to improve patient experience of care.

Keywords

natural language processing, NLP, patient review, patient experience, sentiment analysis, topic analysis

Key Points

1. NLP analysis of large amounts of patient verbatim comments regarding experience can identify specific sentiments and themes for improving experience.
2. The worst experiences are related to communication and relationship.
3. Moderate experience is associated with logistical and process issues.

Introduction

Background

Online review platforms allow patients to describe their experience of medical care.¹ Patients may use online review platforms to learn about clinicians.² These platforms provide information about a clinician’s training and experiences, numerical ratings and average ratings, and verbatim patient comments about their experience. The most negative patient experience comments tend to address aspects of interpersonal communication, trust, and relationship.

For instance, a retrospective study of patient reviews on two online platforms found patients were less satisfied when they perceived the clinician as “uncaring” or “disrespectful.”¹ And another study of 1-star online patient reviews found that “Bedside-manner” was an important factor associated with the lowest patient experience ratings.³ Indeed, the stories that other patients tell about clinicians may be more important to patients than quantitative or objective information such as training, experience, expertise, awards, or quality and safety metrics. For instance, one study of patient selection of a clinician using a simulated rating website containing both subjective and objective

¹ Department of Surgery and Perioperative Care, Dell Medical School, The University of Texas at Austin, Austin, TX, USA

² East Bay Hand Medical Center, San Leandro, CA, USA

Corresponding Author:

David Ring, Dell Medical School, Department of Surgery and Perioperative Care, The University of Texas at Austin, Health Discovery Building Z-stop 0800 1601 Trinity St, Austin, TX 78712, USA.

Email: david.ring@austin.utexas.edu



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

information about potential clinicians found that patients trusted subjective ratings and comments more than objective information regarding clinician experience and training.⁴

Rationale

Studies to date have often used qualitative research methods to identify themes in verbatim patient comments about their experience, a process that requires expertise and is time-consuming. Natural language processing (NLP) can efficiently analyze large amounts of text comments to identify themes and categories embedded in them.⁵ For instance, one study analyzed messages sent through an electronic medical record patient portal using NLP, compared it to the categories found manually with a qualitative approach by medical students and clinicians, and found that NLP identified similar communication theme categories.⁶ NLP was used to identify themes in large sets of comments such as twitter posts⁷ but to our knowledge the use of NLP to identify themes in patient experience comments is a field that is less explored. In one study, sentiment analysis of patient reviews obtained from popular physician review website found that words such as “rude” and “wait” were associated with worse experience and “bedside manner” with better experience.⁸ In another study of patient review gathered from the same platform on family physician and internist doctors, sentiment analysis of patient reviews found that topics such as “caring” were associated with higher experience ratings.⁹ These preliminary studies demonstrate the potential for NLP to help individual clinicians and entire care units to efficiently process large amounts of data about the experience of patients under their care—without the need for specialized training or expertise—in order to help identify areas for personal and organizational improvement.

Questions

To further develop the analysis of themes in patient verbatim comments about their experience with a specific clinician or care unit, we used NLP techniques to analyze a set of physician reviews collected using an online consumer rating program from patients who were less than completely satisfied (ratings of 1 to 4 out of 5) to address the following questions: (1) Which sentiments are associated with specific patient ratings of less than top-rated experience? And (2) What topics are associated with specific numerical ratings?

Method

Data Collection

Approval by the institutional review board was not needed since all the PHI were removed from the data. Ratings of experience from 1 (worst) to 5 (best) along with verbatim text comments describing the experience were available from a commercial enterprise that works with specialty care

practices to help measure and manage patient experience. Among 2188 ratings of 4 or lower, 1117 had associated comments.

Analysis

We analyzed the sample of 1117 ratings of patient ratings of experience ranging from 1 to 4 on a scale from 1 (worst) to 5 (best) that had associated verbatim text comments responding to a prompt asking the patient to suggest opportunities for improvement. First, we performed sentiment analysis and regression on the entire dataset seeking linguistic factors associated with each specific numeric rating. Then, we built topic models to identify themes associated with specific rating values 1 through 4.

Sentiment Analysis. To identify words representing sentiments associated with levels of numerical patient ratings lower than the top score, we used an NLP program (Linguistic Inquiry and Word Count; LIWC-22) to analyze the verbatim patient comments for each rating level.¹⁰ LIWC-22 identifies the proportion of words that correspond to 110 distinct linguistic dimensions and psychological processes. The software also produces eight quantitative variables that represent or summarize subgroups of these dimensions, such as words representing positive or negative emotions. We analyzed the correlation of each numerical experience rating between 1 and 4 (used as the dependent variable) and the 111 LIWC-22 variables (serving as independent variables). Variables with variance inflation factors > 10 or tolerance < 0.1 were considered collinear and were excluded. Correlations with $P < .05$ were advanced into a multivariable linear regression model¹¹ (StataCorp. 2021. *Stata Statistical Software: Release 17*. College Station, TX: StataCorp LLC).

Topic Modeling. To analyze and categorize word clusters (ie, topics) within text comments of ratings 1 to 4 independently, we applied the Meaning Extraction Method (MEM),^{12,13} a NLP technique often referred to as “topic modeling.” MEM identifies the core messages or “themes” presented in language data, by identifying words that often appear together in text. For example, the presence of words like “teacher,” “classroom,” and “student” in a sentence may indicate the topic of “education.” The identification of common word groupings helps reveal the predominant topics in each set of rating-specific comments (Appendix 1)

Results

Which Sentiments are Associated with Patient Ratings of Less Than top-Rated Experience?

Thirty-three of the 111 candidate LIWC-22 variables correlated with ratings of experience between 1 and 4. In multivariable regression, higher ratings of patient experience were associated with positive tone, Positive emotions, Negative

emotions, and Allure and Future focus (Appendix 2, Table 1). Lower experience ratings were associated with higher Word count, Numbers, Negations, Negative tone, Social behavior, Ethnicity, and Motion.

What Topics are Associated with Specific Numerical Ratings?

1-Star Ratings. There were 234 one-star ratings in the sample, the lowest possible rating. Preprocessing yielded 57 unique words with words “doctor” “see” and “time” being the most commonly used (Appendix 2, Table 2). Applying the principal component analysis (Appendix 3) yielded a three-component solution, we identified the following topics to be the most prominent ones: “Listen, concern, collaboration,” “Responsiveness,” and “Wait” (Appendix 2, Table 3; Table 1).

2-Star Ratings. There were 210 two-star ratings in the sample. Preprocessing yielded 48 unique words with words “doctor,” “surgery,” and “patient” most commonly used (Appendix 2, Table 4), and the principal component analysis (Appendix 3) led to a four-component solution. We found these to be the most prominent topics: “Responsiveness,” “Logistics,” “Comfort,” and “Listening, Concern, and Collaboration” (Appendix 2, Table 5; Table 1).

3-Ratings. There were 281 three-star ratings in the sample. Preprocessing yielded 37 unique words with “doctor,” “time,” and “see” being the most common ones (Appendix 2, Table 6), and the principal component analysis (Appendix 3) led to a three-component solution the topics “Wait,” “Listening, Concern, Collaboration,” and “Logistics” were the most prominent (Appendix 2, Table 7; Table 1).

4-Ratings. There were 388 four-star ratings in the sample, the closest rating to complete satisfaction. Preprocessing yielded 26 unique words out of which “doctor,” “time,” and “surgery” were most commonly used (Appendix 2, Table 8), and the principal component analysis (Appendix

3) led to a three-component solution : “Logistics,” “Wait,” and “Pain Alleviation” (Appendix 2, Table 9; Table 1).

Discussion

Qualitative analysis of patient statements regarding experience can identify specific areas of improvement for individual clinicians and larger care units. The process of qualitative research requires resources and expertise that are not readily available. There is precedent for using NLP to analyze a large number of patient statements regarding their experience of care. We measured sentiments and topics in the text comments of patients who were less than fully satisfied with musculoskeletal specialty care (less than 5 out of 5 rating), and—consistent with prior research—found themes of “listening, concern, and collaboration” and “responsiveness” associated with the worst ratings and “logistics” and “Pain” associated with less than completely positive ratings (4 out of 5 stars); and “Wait Times” appeared with both. The ability of NLP to efficiently and effectively identify themes for improvement for clinicians and care teams provides another tool for growth and improvement (Table 2).

Limitations

There are several limitations to our study. First, our dataset is from a limited number of practices. An analysis using data from a larger number of practices might generate different results, although, our findings are relatively consistent across studies done to date, and mostly demonstrate the potential of NLP to be of use. Second, NLP can’t identify indirect language such as sarcasm or slang. Use of a version of NLP specifically designed for analysis of patient comments about experience could be more accurate.

Which Sentiments are Associated with Patient Ratings of Less Than Perfect Experience?

The finding that sentiment analysis discerned differences between levels of ratings of less than perfect experience demonstrates the utility of NLP in measuring and improving patient experience. These findings support a prior study demonstrated that sentiments associated with negative and

Table 1. Representative Words.

Rating	Top 10 unique, non-logistical words
1	Care, knee, wrong, leave, give, MRI, work, listen, experience, recommend
2	Knee, leave, give, problem, MRI, care, experience, result, follow, staff
3	Issue, give, MRI, staff, leave, answer, schedule, question, problem, result
4	Staff, long, knee, great, procedure, rush, leave, give, question, follow

Common terms: Doctor, office, surgery, patient, appointment, visit, call, day
Common theme words: See, time, pain, wait

Table 2. Representative Themes.

	1-star	2-star	3-star	4-star
"Listening, concern, collaboration"	1 st	4 th	2 nd	
"Responsiveness"	2 nd	1 st		
"Wait"	3 rd		1 st	2 nd
"Logistics"		2 nd	3 rd	1 st
"Comfort"		3 rd		
"Pain alleviation"				3 rd

positive experience could be used to generate a numerical score without the ceiling effects typical of current quantitative experience measures.¹⁴ Measures of perceived clinician empathy, communication effectiveness, and willingness to recommend often have a distribution of scores over-represented at the top end of the scale, a sign of loss of data regarding variation. While it might be tempting to assume these ratings represent perfect patient experience, more likely people have suggestions for improvement but are influenced by factors such as social desirability bias to give top scores.¹⁵ Quantitative analysis of patient descriptions of their experience may provide a way to measure experience in a more comprehensive manner less prone to ceiling effects. This has the potential to enhance the ability to track improvements after strategic interventions.

What Topics are Associated with Specific Numerical Ratings?

The observation that the lowest ratings are most closely associated with relationship issues such as listening, concern, and collaboration is consistent with prior studies.^{16,17} For instance, previous studies on 1-star reviews identified that most of the lowest ratings were related to relationship issues (bedside manner, effective communication, perceived empathy, trust) and lengthy wait times.^{16,17} The observation that middle numerical ratings of experience are associated with logistical issues and wait times is in line with a previous study of online reviews from physician profiles on three patient review websites that identified higher average ratings were associated with punctuality and good logistics.¹⁸ The observation that prolonged wait times is seen in both 1-star and 4-star ratings is consistent with a study of patient reviews obtained from a popular online doctor rating website that found a correspondence between longer wait times lower patient satisfaction across all metrics.¹⁹ Another study of thousands of online patient reviews associated “timeliness” with higher experience ratings.²⁰

Conclusion

The ability of NLP to identify themes in large numbers of patient comments is now tested in several venue including portal messages,²¹ group chat sites,⁷ social media posts,²² and patient surveys regarding their experience⁶ as confirmed in the current study. What once might have been a daunting endeavor, to train or hire a person in semi-structured interviews and coding of resultant themes (qualitative research), now has a streamlined alternative in NLP. Testing of the utility of large language models might be the next step, as they can perform many NLP tasks such as identifying relevant themes. Given the consistent identification of difficulties with relationship factors such as trust, perceived empathy, and communication effectiveness at the heart of experiences

that patients rate poorly, individual clinicians and care units can prioritize ensuring that patients feel heard and understood.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors declare that the company from which data is acquired is owned by OF.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ali Azarpey  <https://orcid.org/0000-0002-9915-6288>

David Ring  <https://orcid.org/0000-0002-6506-4879>

Statement of Informed Consent

Informed consent for patient information to be published in this article was not obtained because all the PHI were removed from the data.

Supplemental Material

Supplemental material for this article is available online.

References

1. Orhurhu MS, Salisu B, Sottosanti E, et al. Chronic pain practices: An evaluation of positive and negative online patient reviews. *Pain Phys*. 2019;22(5):E477-86.
2. Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM. Public awareness, perception, and use of online physician rating sites. *JAMA*. 2014;311(7):734-5. doi:10.1001/jama.2013.283194
3. Richman EH, Ogbaudu E, Pollock JR, et al. Characterizing negative online reviews of pediatric orthopaedic surgeons. *J Pediatr Orthop*. 2022;42(5):e533-7. doi:10.1097/BPO.0000000000002121
4. Full article: Choosing a physician on social media: Comments and ratings of users are more important than the qualification of a physician. Accessed March 15, 2024. <https://www-tandfonline-com.ezproxy.lib.utexas.edu/doi/full/10.1080/10447318.2017.1330803>
5. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform*. 2019;125:37-46. doi:10.1016/j.ijmedinf.2019.02.008
6. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int J Med Inf*. 2017;105:110-20. doi:10.1016/j.ijmedinf.2017.06.004
7. Sunkureddi P, Gibson D, Doogan S, Heid J, Benosman S, Park Y. Using self-reported patient experiences to understand patient burden: Learnings from digital patient communities in ankylosing spondylitis. *Adv Ther*. 2018;35(3):424-37. doi:10.1007/s12325-018-0669-1

8. Cho LD, Tang JE, Pitaro N, et al. Sentiment analysis of online patient-written reviews of vascular surgeons. *Ann Vasc Surg.* 2023;88:249-55. doi:10.1016/j.avsg.2022.07.016
9. Nemzer LR, Neymotin F. Concierge care and patient reviews. *Health Econ.* 2020;29(8):913-22. doi:10.1002/hec.4028
10. Boyd RL. The Development and Psychometric Properties of LIWC-22.
11. Mitchell MN. *Interpreting and visualizing regression models using STATA*. Second edition. Stata Press; 2021.
12. Markowitz DM. The meaning extraction method: An approach to evaluate content patterns from large-scale language data. *Front Commun.* 2021;6. Accessed January 26, 2024. <https://www.frontiersin.org/articles/10.3389/fcomm.2021.588823>
13. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *J Res Pers.* 2008;42(1):96-132. doi:10.1016/j.jrp.2007.04.006
14. 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:10.1097/QMH.0000000000000355
15. Badejo MA, Ramtin S, Rossano A, Ring D, Koenig K, Crijns TJ. Does adjusting for social desirability reduce ceiling effects and increase variation of patient-reported experience measures? *J Patient Exp.* 2022;9(4):23743735221079144. doi:10.1177/23743735221079144
16. Richman EH, Kuttner NP, Foster LO, et al. Characterizing single-star negative online reviews of orthopaedic trauma association members. *J Am Acad Orthop Surg.* 2023;31(8):397-404. doi:10.5435/JAAOS-D-22-00631
17. Pollock JR, Arthur JR, Smith JF, et al. The majority of complaints about orthopedic sports surgeons on yelp are nonclinical. *Arthrosc Sports Med Rehabil.* 2021;3(5):e1465-72. doi:10.1016/j.asmr.2021.07.008
18. Kalagara S, Eltorai AEM, DePasse JM, Daniels AH. Predictive factors of positive online patient ratings of spine surgeons. *Spine J.* 2019;19(1):182-5. doi:10.1016/j.spinee.2018.07.024
19. Gupta A, Gupta R, White MD, et al. Patient satisfaction reviews for 967 spine neurosurgeons on healthgrades. *J Neurosurg Spine.* 2022;36(5):869-75. doi:10.3171/2021.8.SPINE21661
20. Yu J, Samuel LT, Yalçın S, Sultan AA, Kamath AF. Patient-recorded physician ratings: What can we learn from 11,527 online reviews of orthopedic surgeons? *J Arthroplasty.* 2020;35(6S):S364-7. doi:10.1016/j.arth.2019.11.021
21. Topaz M, Lai K, Dhopeswarkar N, et al. Clinicians' reports in electronic health records versus patients' concerns in social media: A pilot study of adverse drug reactions of aspirin and atorvastatin. *Drug Saf.* 2016;39(3):241-50. doi:10.1007/s40264-015-0381-x
22. Liu X, Chen H. A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports. *J Biomed Inform.* 2015;58:268-79. doi:10.1016/j.jbi.2015.10.011