Using Natural Language Processing to Automatically Assess Feedback Quality: Findings From 3 Surgical Residencies

Erkin Ötleş, MSE, Daniel E. Kendrick, MD, Quintin P. Solano, Mary Schuller, MSEd, Samantha L. Ahle, MD, MHS, Mickyas H. Eskender, MD, Emily Carnes, and Brian C. George, MD, MAEd

Abstract

Purpose

Learning is markedly improved with high-quality feedback, yet assuring the quality of feedback is difficult to achieve at scale. Natural language processing (NLP) algorithms may be useful in this context as they can automatically classify large volumes of narrative data. However, it is unknown if NLP models can accurately evaluate surgical trainee feedback. This study evaluated which NLP techniques best classify the quality of surgical trainee formative feedback recorded as part of a workplace assessment.

Method

During the 2016–2017 academic year, the SIMPL (Society for Improving Medical Professional Learning) app was used to record operative performance narrative feedback for residents at 3 university-based general surgery residency training programs. Feedback comments were collected for a sample of residents representing all 5 postgraduate year levels and coded for quality. In May 2019, the coded comments were then used to train NLP models to automatically classify the quality of feedback across 4 categories (effective, mediocre, ineffective, or other). Models included support vector machines (SVM), logistic regression, gradient boosted trees, naive Bayes, and random forests. The primary outcome was mean classification accuracy.

Results

The authors manually coded the quality of 600 recorded feedback comments. Those data were used to train NLP models to automatically classify the quality of feedback across 4 categories. The NLP model using an SVM algorithm yielded a maximum mean accuracy of 0.64 (standard deviation, 0.01). When the classification task was modified to distinguish only high-quality vs low-quality feedback, maximum mean accuracy was 0.83, again with SVM.

Conclusions

To the authors' knowledge, this is the first study to examine the use of NLP for classifying feedback quality. SVM NLP models demonstrated the ability to automatically classify the quality of surgical trainee evaluations. Larger training datasets would likely further increase accuracy.

Performance feedback is critical to learning and is highly valued across medical education domains.[1–3] In surgical training, its significance has been established as a powerful means to accelerate improvement in both clinical and technical performance.[4–7] With recent concerns regarding the competence of graduating residents in general surgery,[8,9] an effort has been made by some surgical training programs to standardize the components of quality feedback to improve the assessment of trainee operative performance.[10,11] Unfortunately, it is difficult to ensure that faculty are adhering to these standards when delivering feedback to trainees in practice. Current methods to evaluate feedback quality are labor intensive because they require trained raters to review and classify the quality of individual recorded feedback.[12,13] This is a growing problem with the widespread adoption of smartphone assessment applications, which have greatly increased the volume of narrative feedback available to trainees.[14–16] Alternative methodologies for feedback quality assurance are therefore needed to efficiently identify instructors who are not meeting standards and who might benefit most from targeted faculty development.

Emerging technology in machine learning (ML) may be an automated solution. A subfield of ML, known as natural language processing (NLP), includes algorithms developed for automated text analysis. NLP has been successfully developed in other fields to classify document sentiment, identify important entities in text, and even automatically translate text from one language to another. In medical education, a variety of NLP techniques have been used to automate the evaluation of trainee documentation and clinical experiences.[17–19] However, to our knowledge, NLP techniques have never been used to assess the quality of feedback provided to trainees by faculty.[20–22] In an effort to better understand how automated feedback quality assurance could be implemented using NLP, we investigated the accuracy of different NLP models to classify the quality of feedback provided to surgical trainees.

Method

Study population

We conducted this analysis in May 2019 at the University of Michigan Medical School. Data were collected from a convenience sample of 3 university-based general surgery residency training programs, all part of large

Please see the end of this article for information about the authors.

Correspondence should be addressed to Quintin P. Solano, University of Michigan Medical School, 1301 Catherine St., Ann Arbor, MI 48109; telephone: (313) 433-2928; email: qsolano@med.umich.edu.

Acad Med. 2021;96:1457–1460.
First published online May 4, 2021
doi: 10.1097/ACM.0000000000004153
Copyright © 2021 by the Association of American Medical Colleges

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/B110.

university-affiliated academic institutions (Northwestern Feinberg School of Medicine, Massachusetts General Hospital/Harvard, Southern Illinois University). These training programs are members of the Society for Improving Medical Professional Learning (SIMPL), a not-for-profit educational research consortium.[23] Data analyzed in this study were originally collected as part of a related study aimed at characterizing the overall quality of narrative feedback data using human raters.[13] Data were collected for trainees from all 5 postgraduate years who were residents during the 2016–2017 academic year.

Feedback instruments

End of rotation evaluation narrative items. The end of rotation (EOR) evaluation is both a summative and formative instrument intended to be completed by faculty within 1 week of the resident completing a rotation on their service. The EOR evaluation at each institution varied slightly, but the items we examined were only those free-text items that allowed faculty to provide trainees narrative feedback on their operative performance.

SIMPL narrative items. SIMPL is a quality improvement collaborative that developed and maintains the SIMPL app, a smartphone-based workplace assessment tool.[23] The SIMPL app is a reference implementation of an evidence-based guideline for operative performance assessment.[11] The app asks raters to evaluate directly observed operative performance within 72 hours of the completion of an operation and answers 3 questions regarding the observed resident operative performance, resident autonomy, and the patient-specific case complexity. Faculty can also dictate formative feedback for the resident about that specific case. We analyzed these narrative feedback. For further background about the SIMPL app including specific evaluator instructions, ratings instruments, resident usage, and other characteristics, please see George and colleagues.[24]

Data collection

We retrospectively collected EOR and SIMPL app narrative evaluation data of the subject residents. EOR data were collected directly from the 3 residency programs and SIMPL app data were obtained from the SIMPL consortium. We obtained exemption from the University of Michigan's institutional review board as this was a retrospective secondary data analysis.

We began by de-identifying and collecting EOR evaluation narrative comments. The audio from recorded SIMPL dictations was separately transcribed using Google Cloud Platform Speech-to-Text (Google LLC, Mountain View, California) and transcription errors corrected by a third author (E.C.) who did not participate in subsequent quality coding. Comments were randomly but separately sampled from each data source to ensure a balanced dataset.

Quality coding system

Following the methods outlined in Ahle and colleagues,[13] comments were coded by surgeon raters (S.A., M.E.) blinded to the data source, institution, and resident postgraduate year, as being specific or general, encouraging or not encouraging, and corrective or not corrective. The unit of analysis was the entire comment. To qualify for a particular category (e.g., specific versus general), only 1 portion of a recorded feedback narrative needed to meet descriptions for that category. We used an additional category of "irrelevant" for rater comments that did not include any feedback on the resident's operative performance.

We split comments into groups and rated them in phases. After each phase, both coders met to identify discrepancies, explore the sources of ratings variation, and refine their coding scheme for subsequent coding phases. When needed, the coders shared examples of categorization inconsistencies with the larger interdisciplinary research team for further discussion. After all comments were coded, the 2 coders reviewed each comment with discordant categories and agreed upon a final consensus category. Using a modification from the previously defined scoring rubric,[13] we then used the categories to classify the feedback as effective (E), mediocre (M), ineffective (I), or other (O). These were the codes used in our analysis. For examples of feedback and their associated ratings, see Supplemental Digital Appendix 1, at http://links.lww.com/ACADMED/B110.

Statistical analysis

We analyzed all of the narrative data along with its associated quality codes separately with multiple NLP pipelines, composed of several different ML models. These ML models were random forest (RF), naive Bayes (NB), gradient boosted trees (GB), logistic regression (LR), and support vector machines (SVM). We constructed all of these NLP pipelines with the Python packages NumPy (Trelgol Publishing, 2006) and SKLearn (Scikit-learn, 2011). We chose these specific ML models because of their interpretability, efficient implementations, and widespread adoption in the ML community. Briefly, the NLP pipeline process began with extracting tokens (single words or strings of multiple words) from raw evaluation text. Each evaluation was turned into a series of vectors, which represented the counts of tokens present in the raw evaluation text. These vectors were then fed to the individual ML models mentioned above. Pipeline training was conducted using cross-validation. Details regarding models and the training procedure are presented in Supplemental Digital Appendix 2, at http://links.lww.com/ACADMED/B110.

For our primary analysis, the primary outcome was the 4-category classification performance for each NLP model, which we assessed using 5-fold cross-validation. We also examined the classification accuracy of our models when used to make a binary classification of operative performance feedback as high vs low quality. To perform this analysis, we grouped the 4 quality codes into 2 categories and compared those results with the original primary analysis results. Specifically, we compared the classification accuracy when using the narrative data recoded as {E,M} (high quality) vs {I, O} (low quality) compared with the original 4-category quality coding (any of {E, M, I, O}).

Results

We collected 600 training instances, 300 from the EOR evaluations and 300 from SIMPL. Of these, 207 (34.5%) were rated as E, 110 (18.3%) as M, 198 (33.0%) as I, and 85 (14.2%) as O. For further characterization of this dataset, please see Ahle and colleagues.[13]

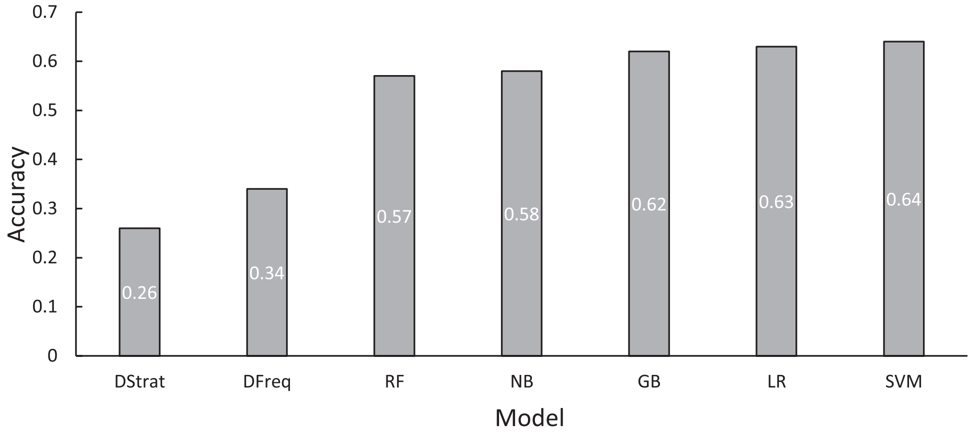

Figure 1 Accuracies of NLP models when classifying SIMPL narrative transcripts from the SIMPL dataset and end of rotation feedback into E, M, I, O classifications, from a study of automatic assessment of feedback quality to surgical residents at 3 institutions, 2016–2017. In total, there were 600 items: 300 SIMPL transcripts and 300 end of rotation narratives. DStrat and DFreq represent the 2 naive baseline models and the other models reported are their own packages with their own respective, unique algorithm. Abbreviations: NLP, natural language processing; SIMPL, Society for Improving Medical Professional Learning app; E, effective; M, mediocre; I, ineffective; O, other; DStrat, dummy stratified; DFreq, dummy frequency; RF, random forest; NB, naive Bayes; GB, gradient boosted trees; LR, logistic regression; SVM, support vector machines.

Four-category classification

When comparing NLP models for the task of classification performance, the NLP pipeline with an SVM model achieved the highest mean accuracy of 0.64 (standard deviation [SD], 0.01). Other models achieved mean accuracies of 0.26 to 0.62. Figure 1 shows a comparison of the mean performance for all models.

Binary classification

When the classification task was simplified to categorize feedback as high quality {E, M} vs low quality {I, O}, SVM remained the most performant model (mean accuracy, 0.83; SD, 0.01). LR (mean accuracy; 0.81; SD, 0.04) was the second most performant model. Results for all the models are summarized in Table 1.

Table 1

Accuracy Performance of Natural Language Processing Models on Grouped Quality Ratings Using Original and Binary Coding, From a Study of Automatic Assessment of Feedback Quality to Surgical Residents at 3 Institutions, 2016–2017

	Mean accuracy (SD)	
Model	Original coding: E, M, I, O	Binary coding: {E, M} vs {I, O}
DStrat	0.25 (0.07)	0.55 (0.02)
DFreq	0.34 (0.00)	0.52 (0.00)
RF	0.57 (0.02)	0.77 (0.02)
NB	0.57 (0.03)	0.79 (0.02)
GB	0.61 (0.02)	0.83 (0.02)
LR	0.62 (0.01)	0.81 (0.01)
SVM	0.64 (0.01)	0.83 (0.01)

Abbreviations: SD, standard deviation; E, effective; M, mediocre; I, ineffective; O, other; DStrat, dummy stratified; DFreq, dummy frequency; RF, random forest; NB, naive Bayes; GB, gradient boosted trees; LR, logistic regression; SVM, support vector machines.

Discussion

NLP models demonstrated the ability to automatically classify the quality of surgical trainee feedback from faculty. These findings mark what we believe is the first attempt to automate feedback quality classification. When compared with human-coded quality classifications, 4-category NLP classification achieved a peak accuracy of 0.64 when using an SVM model. Improvements in classification accuracy were noted by simplifying the classification task. Specifically, when NLP models were trained to categorize feedback as high vs low quality, SVM achieved accuracies of 0.83. Ultimately, the decision to classify feedback quality into 2 vs 4 categories should be driven not by performance but by the desired use-case of the model.

Implementation

While the dataset that we used for this pilot study is relatively small, our findings highlight both the promise and limitations of NLP.[22,25] NLP models such as those described above could be used to automatically classify the quality of the many thousands of comments recorded by faculty and collected each year by residency programs. However, further refinements will be needed before this technology can be used to reliably classify single comments. In other words, our NLP system cannot be used to immediately assess the quality of feedback provided to trainees in real time. Instead, fair assessment of feedback quality for individual faculty will likely require analysis of multiple comments. To that end, we do not envision programs reporting the categorizations of individual feedback items to faculty. Instead, these NLP models could be used to flag assessors who repeatedly generate low-quality feedback. Programs could then help allocate faculty development resources to these assessors to help raise the overall quality of their feedback to residents. These resources could be included in an automated email that highlights the rater’s feedback quality, points faculty to existing or new curricula, and provides action items for improvement. Aggregate statistics of feedback quality could also serve as an outcome measure for local educational quality improvement efforts.

Limitations

All data analyzed in this study were from 3 large academic surgical training centers and may not be representative of the general population of evaluations on surgical trainees. Furthermore, at the time of analysis, the data were already 2 years old, although we do not expect the time of data collection to affect how well NLP might classify feedback quality. We also did not assess the impact of accents, which could be significant and might falsely diminish the rated quality leading to biased assessment of non-native English speakers. This will need to be addressed if such a system was to be implemented in practice. Furthermore, the evaluations were all coded by surgeons based on text transcribed from Google Cloud Platform Speech-to-Text. While capable, machine transcription is still fallible and produces transcription errors that are interpretable to humans but may adversely affect NLP models. This could be remedied by improvements in automated transcription or having humans audit the transcriptions, although current NLP technologies may be able to overcome these issues given sufficient training data. Finally, our study does not have data from other evaluation systems. Our models may therefore not be generalizable to different evaluation instruments and processes.

Conclusions

NLP models trained on surgical resident evaluations were able to automatically classify the quality of operative performance feedback. Accuracy is affected by how feedback quality is categorized. With additional training data, NLP models may be able to achieve higher levels of accuracy. These tools could also be integrated into workplace-based assessment applications. Doing so would allow faculty to receive automated feedback about the quality of the feedback they are providing to trainees. It would also feasibly permit programs to target faculty development resources specifically to those who would most benefit. In an era of finite operative experiences for trainees, these types of interventions could improve operative teaching and expand the impact for learners. Maximizing the value of every operative experience in training is critical to ensure safe patient care when trainees enter independent practice.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: Deemed exempt by the University of Michigan institutional review board.

Previous presentations: Accepted for poster presentation at the Association for Surgical Education (ASE) Annual Meeting; April 28, 2020; Seattle, Washington (canceled due to the COVID-19 pandemic; a virtual abstract was presented via Twitter).

Data: End of rotation feedback and transcript data obtained from and with permission of the Society for Improving Medical Professional Learning (SIMPL).

E. Ötleş is Medical Scientist Training Program fellow, Department of Industrial and Operations Engineering, University of Michigan Medical School, Ann Arbor, Michigan.

D.E. Kendrick is assistant professor, Department of Surgery, University of Minnesota Medical School, Minneapolis, Minnesota.

Q.P. Solano is a third-year medical student, University of Michigan Medical School, Ann Arbor, Michigan.

M. Schuller is senior project manager, Department of Surgery, University of Michigan Medical School, Ann Arbor, Michigan.

S.L. Ahle is a resident, Department of Surgery, Yale School of Medicine, New Haven, Connecticut.

M.H. Eskender is a resident, Department of Surgery, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

E. Carnes is research assistant, Department of Surgery, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

B.C. George is assistant professor and director, Center for Surgical Training and Research, Department of Surgery, University of Michigan Medical School, Ann Arbor, Michigan.

References

1 Wolverton SE, Bosworth MF. A survey of resident perceptions of effective teaching behaviors. Fam Med. 1985;17:106–108.

2 Heckman-Stone C. Trainee preferences for feedback and evaluation in clinical supervision. Clin Supervisor. 2004;22:21–33.

3 Taylor DC, Hamdy H. Adult learning theories: Implications for learning and teaching in medical education: AMEE guide no. 83. Med Teach. 2013;35:e1561–e1572.

4 Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. Surg Endosc. 2007;21:2240–2243.

5 Boyle E, Al-Akash M, Gallagher AG, Traynor O, Hill AD, Neary PC. Optimising surgical training: Use of feedback to reduce errors during a simulated surgical procedure. Postgrad Med J. 2011;87:524–528.

6 Strandbygaard J, Bjerrum F, Maagaard M, et al. Instructor feedback versus no instructor feedback on performance in a laparoscopic virtual reality simulator: A randomized trial. Ann Surg. 2013;257:839–844.

7 Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M. The impact of feedback of intraoperative technical performance in surgery: A systematic review. BMJ Open. 2015;5:e006759.

8 Mattar SG, Alseidi AA, Jones DB, et al. General surgery residency inadequately prepares trainees for fellowship: Results of a survey of fellowship program directors. Ann Surg. 2013;258:440–449.

9 George BC, Bohnen JD, Williams RG, et al; Procedural Learning and Safety Collaborative (PLSC). Readiness of US general surgery residents for independent practice. Ann Surg. 2017;266:582–594.

10 van de Ridder JM, Stokking KM, McGaghie WC, ten Cate OT. What is feedback in clinical education? Med Educ. 2008;42:189–197.

11 Williams RG, Kim MJ, Dunnington GL. Practice guidelines for operative performance assessments. Ann Surg. 2016;264:934–948.

12 Ali A, Bussey M, O'Flynn K, Eardley I. Quality of feedback using workplace based assessments in urological training. British J Med Surg Urology. 2012;5:39–43.

13 Ahle SL, Eskender M, Schuller M, et al. The quality of operative performance narrative feedback: A retrospective data comparison between end of rotation evaluations and workplace-based assessments [published online ahead of print June 4, 2020]. Ann of Surg. doi:10.1097/SLA.0000000000003907.

14 Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. Surgery. 2005;138:640–647.

15 Wohaibi EM, Earle DB, Ansanitis FE, Wait RB, Fernandez G, Seymour NE. A new web-based operative skills assessment tool effectively tracks progression in surgical resident performance. J Surg Educ. 2007;64:333–341.

16 Bohnen JD, George BC, Williams RG, et al; Procedural Learning and Safety Collaborative (PLSC). The feasibility of real-time intraoperative performance assessment with SIMPL (System for Improving and Measuring Procedural Learning): Early experience from a multi-institutional trial. J Surg Educ. 2016;73:e118–e130.

17 Chary M, Parikh S, Manini AF, Boyer EW, Radeos M. A review of natural language processing in medical education. West J Emerg Med. 2019;20:78–86.

18 Hasan Sapci A, Aylin Sapci H. Artificial intelligence education and tools for medical and health informatics students: Systematic review. JMIR Med Educ. 2020;6:e19285.

19 Denny JC, Bastarache L, Sastre EA, Spickard A 3rd. Tracking medical students' clinical experiences using natural language processing. J Biomed Inform. 2009;42:781–789.

20 Goldberg Y. Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies. 2017;1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037. Accessed March 22, 2021.

21 Martin JH, Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, NJ: Pearson/Prentice Hall; 2009.

22 Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Comput Intell. 2018;13:55–75.

23 SIMPL. About the SIMPL collaborative. http://www.simpl.org. Published 2020. Accessed March 22, 2021.

24 George BC, Bohnen JD, Schuller MC, Fryer JP. Using smartphones for trainee performance assessment: A SIMPL case study. Surgery. 2020;167:903–906.

25 Banko M, Brill E. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. Proceedings of the First International Conference on Human Language Technology Research. March 2001.