Exploring the Use of Natural Language Processing to Understand Emotions of Trainees and Faculty Regarding Entrustable Professional Activity Assessments

Devin Johnson, PhD
Sonaina Chopra, BA
Elif Bilgic, PhD

ABSTRACT

Background In medical education, artificial intelligence techniques such as natural language processing (NLP) are starting to be used to capture and analyze emotions through written text.

Objective To explore the application of NLP techniques to understand resident and faculty emotions related to entrustable professional activity (EPA) assessments.

Methods Open-ended text data from a survey on emotions toward EPA assessments were analyzed. Respondents were residents and faculty from pediatrics (Peds), general surgery (GS), and emergency medicine (EM), recruited for a larger emotions study in 2023. Participants wrote about their emotions related to receiving/completing EPA assessments. We analyzed the frequency of words rated as positive via a validated sentiment lexicon used in NLP studies. Specifically, we were interested if the count of positive words varied as a function of group membership (faculty, resident), specialty (Peds, GS, EM), gender (man, woman, nonbinary), or visible minority status (yes, no, omit).

Results A total of 66 text responses (30 faculty, 36 residents) contained text data useful for sentiment analysis. We analyzed the difference in the count of words categorized as positive across group, specialty, gender, and being a visible minority. Specialty was the only category revealing significant differences via a bootstrapped Poisson regression model with GS responses containing fewer positive words than EM responses.

Conclusions By analyzing text data to understand emotions of residents and faculty through an NLP approach, we identified differences in EPA assessment-related emotions of residents versus faculty, and differences across specialties.

Introduction

Entrustable professional activities (EPAs) are essential for residents to achieve competence to ensure effective patient care at the end of residency.[1,2] Based on initial research targeting EPA assessments, preliminary findings show that the overall perception of trainees regarding EPA assessments is mostly negative, with trainees feeling overwhelmed, anxious, and frustrated, mostly due to logistical difficulties and an uncertainty of the educational benefit.[3-7]

Data from these studies are often qualitative in nature, typically captured through open-ended text responses in surveys, which require rigorous and methodical qualitative analyses to glean insights from this data. Natural language processing (NLP) methods provide a useful tool for supporting qualitative analysis of text data, especially when text data is vast or potential trends or patterns would need to be understood before conducting in-depth qualitative analysis. One method within NLP, sentiment analysis (SA), allows for the emotional valence of text data to be assessed, typically in the form of validated dictionaries or lexicons that categorize words by specific emotions or provide numerical ratings for words regarding polarity.

When used appropriately, SA offers the opportunity to assess whether written feedback is positive or negative, which could play an important role in medical education.[8] For example, negative emotions are shown to hinder performance; yet, initial research targeting EPA assessments shows that the overall emotions of trainees are mostly negative.[3] Therefore, at the program evaluation level, programs could explore whether trainees and faculty have positive and/or negative emotions regarding EPA assessments, and if negative, where the negative emotions stem from, to develop initiatives to create an optimized EPA assessment process.

As current guidance on the use of NLP in the medical education space is focused on project-specific approaches,[8] in this article, we showcase an example of SA when applied to EPA assessment feedback of

DOI: http://dx.doi.org/10.4300/JGME-D-23-00526.1

residents and faculty. Here, we show how researchers and educators interested in conducting SA on written feedback can convert text responses into numerical data and perform statistical analyses to analyze potential trends in emotional valence across categories of interest.

Methods

To explore the applicability of SA in a medical education context, we use written feedback from residents and faculty as a case study. Specifically, we focus on the emotions that residents and faculty express with regard to receiving/completing EPA assessments. This data comes from a survey of residents and faculty from pediatrics (Peds), general surgery (GS), and emergency medicine (EM), conducted in 2023. These 3 specialties were chosen in the original study (not published yet) in order to understand EPA assessment-related emotions across different specialties and years of EPA assessment implementation (Peds in 2021, GS in 2020, EM in 2018), in addition to being a convenience sample of the study teams' own specialties. Participants were from McMaster University in Hamilton, Ontario, Canada, which is an urban program with regional and main campuses and has postgraduate medical education programs including family medicine and 57 specialty and subspecialty programs. First, to help participants start reflecting on their emotions related to EPA assessment experiences, they completed the Medical Emotion Scale (MES), which is a self-reporting standardized questionnaire that measures the intensity of 20 unique emotions on a 5-point Likert scale.[9] Then, participants answered the following open-ended question: "In the space below, please explain the factors that may have contributed to your answers above [responses to MES], regarding your current feelings about EPA assessments." Additionally, demographics questions were asked, including group membership (faculty, residents), specialty (Peds, GS, EM), gender (man, woman, nonbinary), or visible minority status (yes, no, omit).

The principal analyst for the project examined the frequency of words categorized as positive by the National Research Council Canada (NRC) Emotion Lexicon,[10] a validated lexicon,[11] accessed via the tidytext package.[12] To convert the raw text data into quantifiable information, we first converted the data into long format (where each row is a word from each respondent's text data) and mapped our sentiment lexicon (containing words and their categorized emotions) onto the long format data frame. From there, we counted the number of unique emotion words per participant before reverting the data frame into a wide format (with each row representing a participant's text data and the count of words related to each emotion).

The tidytext framework in the programming language R facilitates such analyses at no cost, given that both R and the package are open source and maintained by the developers. Analysts familiar with data manipulation, including the conversion of data from wide to long and vice versa, can leverage the functions from these packages to convert text data into a quantified format. For a visual representation of the raw text data, the code used in the transformation, and the final quantified version of the text data, please refer to FIGURE 1.

Given the data available to us, we explored to what extent the count of words categorized as positive varied as a function of group membership or specialty. We placed emphasis on positive words as the key emotion, as positive words were most captured in our dataset, and there was a lower proportion of words across the other emotions in the NRC lexicon captured in our dataset. We also analyzed differences in positive word count by demographic factors such as gender and visible minority status. For our analytic strategy, we applied Poisson regression models, a useful method for analyzing language data across conditions.[13] Due to our small sample size, we conducted bootsrapped versions of each model with 10 000 iterations to assess the robustness of our findings. Ethics approval was received from the Hamilton Integrated Research Ethics Board.

Results

Out of the 91 respondents from the original survey, 73 provided a response to our open-ended question. Out of this 73, 66 respondents (30 faculty, 36 residents) had usable text data that could reliably be mapped onto the sentiment lexicon used for this case study. Analyzing positive word frequency among specialties (FIGURE 2) revealed that relative to EM, GS responses had fewer positive words (b=-1.26; z(63)=-4.52; $P<.001$; 95% CI -1.84, -0.74). There was no difference in the count of positive words for Peds relative to EM (P=.81). Our bootstrapped model revealed similar results in terms of statistical significance, with GS reporting fewer positive words than EM (P=.01), and no difference in the count of positive words between Peds and EM (P=.89).

Results from a bootstrapped Poisson regression model revealed a nonsignificant effect of group membership (faculty compared to residents) on the frequency of positive words (P=.24). For demographics, subsequent bootstrapped models revealed no effect of gender (P=.12) or minority status (P=.89) on positive word frequency.

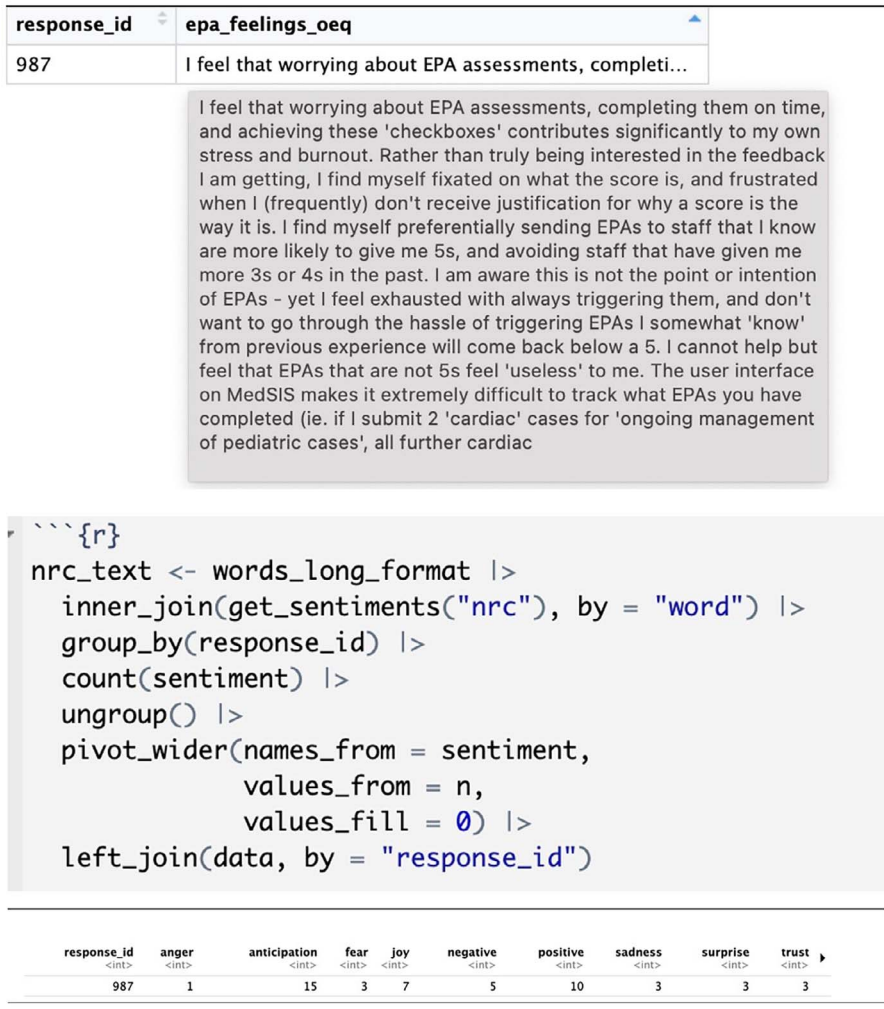

FIGURE 1
Visual Representation of Raw Text Data, Code Used in the Transformation, and Final Quantified Version of Text Data

Discussion

Our SA found differences in the counts of positive words on EPA assessments, particularly by specialty. Such findings add support for the use of NLP (SA in particular) within medical education research. For example, recent work has shown sentiment lexicons such as the NRC can feasibly detect meaningful differences in emotional valence of feedback for trainees.[14] Such lexicons, when leveraged with statistical methods as those used in this case study, also provide useful analyses in open-ended text data trends.[13]

While the use of these methods holds promise, limitations do exist. First, such methods analyze text only by individual words, limiting the full context to be captured.[15] Second, lexicons such as the NRC, while useful, may not capture all relevant aspects of text data within the medical education context.[8] We argue that for this limitation, the content knowledge of researchers can be crucial in conducting an impactful analysis. Researchers with extensive institutional and context-specific knowledge can be useful in the development of a context-specific lexicon used to analyze data from their study.[14] Using our case study as an example, if medical education researchers are aware of specific words of interest pertaining to feedback or assessment, then a quantitative analysis of the frequency of those words can be conducted by mapping that custom lexicon onto text data. As such, contextual knowledge makes fruitful collaboration not only useful but necessary to leverage NLP methods in the field.

With our NLP analysis, since certain differences (both statistically significant and nonsignificant) have been observed for residents and faculty as well as across specialties and gender, we will use these results as a starting point. Specifically, a qualitative study (eg, data collection through interviews) will be conducted with thematic analysis[16] to further understand why residents and faculty feel certain ways across

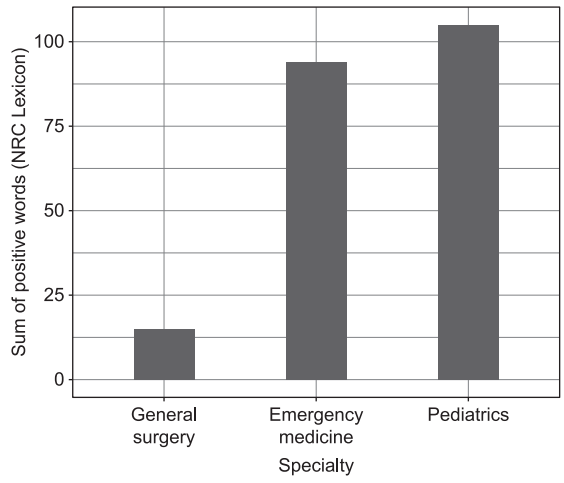

FIGURE 2
Overall Count of Positive Words by Specialty
Abbreviation: NRC, National Research Council Canada.

specialties and other identity factors. For example, based on our findings, one interview question could focus on exploring why there are EPA assessment-related emotion differences across specialties.

Conclusions

By analyzing text data to understand emotions through an NLP approach, we identified differences in EPA assessment-related emotions of residents versus faculty, and differences across specialties.

References

1. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach*. 2010;32(8):676-682. doi:10.3109/0142159X.2010.500704
2. Sherbino J, Regehr G, Dore K, Ginsburg S. Tensions in describing competency-based medical education: a study of Canadian key opinion leaders. *Adv Health Sci Educ Theory Pract*. 2021;26(4):1277-1289. doi:10.1007/s10459-021-10049-8
3. Bilgic E, Turkdogan S, Harley JM. Entrustable professional activity assessments in surgery: competing perspectives, practices, expectations, and future roles? *Glob Surg Educ*. 2023;2:22. doi:10.1007/s44186-022-00099-w
4. Fédération des médecins résidents du Québec. Year 3 of implementation of competence by design: negative impact still outweighs theoretical benefits. Published July 2020. Accessed April 3, 2024. https://fmrq.qc.ca/wp-content/uploads/2022/07/fmrq-report-cbd-implementation-year-3_1.pdf
5. Dagnone JD, Bandiera G, Harris K. Re-examining the value proposition for competency-based medical education. *Can Med Educ J*. 2021;12(3):155-158. doi:10.36834/cmej.68245
6. David V, Walsh M, Lockyer J, Mintz M. Entrustable professional activities: an analysis of faculty time, trainee perspectives and actionability. *Can J Gen Intern Med*. 2021;16(1):8-13. doi:10.22374/cjgim.v16i1.415
7. Royal College of Physicians and Surgeons of Canada. Competence by design: resident pulse check report 2022. Accessed April 2, 2024. https://www.royalcollege.ca/content/dam/documents/accreditation/competence-by-design/non-resource-documents/infographic-e.pdf
8. Costa-Dookhan KA, Maslej MM, Donner K, Islam F, Sockalingam S, Thakur A. Twelve tips for natural language processing in medical education program evaluation [published online ahead of print February 19, 2024]. *Med Teach*. doi:10.1080/0142159X.2024.2316223
9. Duffy MC, Lajoie SP, Pekrun R, Lachapelle K. Emotions in medical education: examining the validity of the Medical Emotion Scale (MES) across authentic medical learning environments. *Learn Instruct*. 2020;70:101150. doi:10.1016/j.learninstruc.2018.07.001
10. Mohammad SM, Turney PD. NRC emotion lexicon. National Research Council Canada. Published November 15, 2013. Accessed April 2, 2024. https://nrc-publications.canada.ca/eng/view/ft/?id=0b6a5b58-a656-49d3-ab3e-252050a7a88c
11. Borchers C, Rosenberg J, Gibbons B, Burchfield M, Fischer C. To scale or not to scale: comparing popular sentiment analysis dictionaries on educational Twitter data. Paper presented at: Fourteenth International Conference on Educational Data Mining (EDM 2021); July 2021; Paris, France. Accessed April 2, 2024. https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_122.pdf
12. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. *J Open Source Software*. 2016;1(3):37. doi:10.21105/joss.00037
13. Winter B, Bürkner P-C. Poisson regression for linguists: a tutorial introduction to modelling count data with brms. *Language Linguistics Compass*. 2021;e12439. doi:10.1111/lnc3.12439
14. Lu KJQ, Meaney C, Guo E, Leung FH. Evaluating the applicability of existing lexicon-based sentiment analysis techniques on family medicine resident feedback field notes: retrospective cohort study. *JMIR Med Educ*. 2023;9:e41953. doi:10.2196/41953
15. Juluru K, Shih H-H, Murthy KNK, Elnajjar P. Bag-of-words technique in natural language processing: a primer for radiologists. *RadioGraphics*. 2021;41(5):1420-1426. doi:10.1148/rg.2021210025

16. Braun V, Clarke V. Reflecting on reflexive thematic analysis. *Qual Res Sport Exercise Health*. 2019; 11(4):589-597. doi:10.1080/2159676x.2019.1628806

Devin Johnson, PhD, is a recent graduate of the PhD program, Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, Ontario, Canada; **Sonaina Chopra, BA,** is a Research Assistant, McMaster Education Research, Innovation and Theory (MERIT) Program, McMaster University, Hamilton, Ontario, Canada; and **Elif Bilgic, PhD,** is an Assistant Professor and Education Scientist, Department of Pediatrics and McMaster Education Research, Innovation and Theory (MERIT) Program, McMaster University, Hamilton, Ontario, Canada.

Funding: This work was supported in part by the Social Sciences and Humanities Research Council of Canada.

Conflict of interest: The authors declare they have no competing interests.

The authors would like to thank Amy Keuhl for her help during manuscript writing.

Corresponding author: Elif Bilgic, PhD, McMaster University, Hamilton, Ontario, Canada, bilgice@mcmaster.ca, X @ElifBilgic16

Received July 26, 2023; revisions received December 6, 2023, and March 18, 2024; accepted March 25, 2024.