Evaluation & the Health Professions
2023, Vol. 46(3) 225–232

DOI: 10.1177/01632787231158128
journals.sagepub.com/home/ehp


Natural Language Processing of Learners' Evaluations of Attendings to Identify Professionalism Lapses

Janae K. Heath[1], Caitlin B. Clancy[1], William Pluta[1], Gary E. Weissman[1,2,3], Ursula Anderson[1], Jennifer R. Kogan[1], C. Jessica Dine[1], and Judy A. Shea[1,2]

Abstract

Unprofessional faculty behaviors negatively impact the well-being of trainees yet are infrequently reported through established reporting systems. Manual review of narrative faculty evaluations provides an additional avenue for identifying unprofessional behavior but is time- and resource-intensive, and therefore of limited value for identifying and remediating faculty with professionalism concerns. Natural language processing (NLP) techniques may provide a mechanism for streamlining manual review processes to identify faculty professionalism lapses. In this retrospective cohort study of 15,432 narrative evaluations of medical faculty by medical trainees, we identified professionalism lapses using automated analysis of the text of faculty evaluations. We used multiple NLP approaches to develop and validate several classification models, which were evaluated primarily based on the positive predictive value (PPV) and secondarily by their calibration. A NLP-model using sentiment analysis (quantifying subjectivity of the text) in combination with key words (using the ensemble technique) had the best performance overall with a PPV of 49% (CI 38%-59%). These findings highlight how NLP can be used to screen narrative evaluations of faculty to identify unprofessional faculty behaviors. Incorporation of NLP into faculty review workflows enables a more focused manual review of comments, providing a supplemental mechanism to identify faculty professionalism lapses.

Keywords

professionalism, medical education, faculty professionalism, natural language processing, identification of professionalism lapses

Faculty mistreatment of trainees is pervasive and destructive, leading to serious negative impacts on learning outcomes and trainee well-being (Cook et al., 2014; Hu et al., 2019; Karnieli-Miller et al., 2010; Oser et al., 2014; Richman et al., 1992; Wilkinson et al., 2006; 2009). Fostering a safe, positive learning environment and continually displaying respectful interactions with patients and colleagues are essential elements of faculty professionalism (Chung et al., 2018). Lapses in professional behavior can directly or indirectly result in trainee mistreatment, occurring along a spectrum from blatant mistreatment (e.g., verbal or physical abuse, sexual harassment) to more subtle mistreatment (e.g., neglect of teaching duties, disrespectful behavior, favoritism, inappropriate workload and/or lack of supervision) (Gan & Snell, 2014). Providing feedback and remediation to faculty is a critical step in addressing learner mistreatment. However, identification of individual faculty members that exhibit unprofessional behavior is challenging (Chung et al., 2018; Ross et al., 2018). In fact, the proportion of trainees who experience or witness mistreatment significantly exceeds the proportion of trainees who ultimately report these behaviors, with nearly 80% of respondents who experienced mistreatment not reporting based on national data (AAMC Graduation Questionnaire GQ, 2021).

Reporting is limited by multiple factors, including a perception that the behavior was not serious enough to warrant reporting, the uncertainty with and burdensome nature of reporting processes, fear of reprisal, and a sense that no action would be taken (Chung et al., 2018). Strategies to increase reporting have included

[1]Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[2]Leonard Davis Institute of Health Economics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[3]Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

Corresponding Author:
Janae K. Heath, Department of Medicine, University of Pennsylvania Perelman School of Medicine, 3600 Spruce Street, 822 West Gates Building, Philadelphia, PA 19104, USA.
Email: Janae.heath@pennmedicine.upenn.edu

clarification of professionalism standards, education on reporting mechanisms, and implementation of confidential electronic reporting systems (Mazer et al., 2018; Ross et al., 2018). Despite these strategies, significant gaps remain and formal reporting pathways undercount faculty professionalism lapses (AAMC Graduation Questionnaire GQ, 2021).

An additional source for capturing unprofessional behavior is the narrative evaluations of faculty (Chung et al., 2018), where learners may reference unprofessional behavior that they did not label as mistreatment and/or did not meet their threshold for formal reporting. However, reviewing narrative evaluations of faculty requires manual review of large numbers of evaluations. This manual process is both resource- and time-intensive, and ultimately limits the ability for timely interventions. Specifically, within our institution, review of 500-700 comments for professionalism lapses by an experienced reviewer requires approximately 1 hour of time.

One potential strategy to augment manual review is to use computer-based language analysis, or natural language processing (NLP). Natural language processing techniques employ computerized algorithms to interpret large quantities of text data, translating narrative text into structured forms for quantitative evaluation. Applying this automated and computerized approach to data provides the ability to create classification models that learn to identify patterns in the training data, which can then be applied to identify those same themes and patterns in future text samples. In cases of large quantities of narrative evaluation, NLP offers a mechanism to rapidly screen large numbers of evaluations for patterns and identify a subset for further review.

NLP techniques have been applied in medical education research across the continuum (Chary et al., 2019; French et al., 2014; Gierl et al., 2014; Heath & Dine, 2019; Tremblay et al., 2019). Specifically, NLP procedures have been used to quantify differential word use in faculty evaluations (looking for gender differences), identify the words used in struggling resident evaluations, standardize essay scoring for medical trainees, and standardize competency committee discussions. However, the use of NLP techniques to identify faculty professionalism lapses in narrative evaluations has not been described. NLP has the potential to streamline and augment current manual review processes and may ultimately identify professionalism lapses that are not currently being captured.

Therefore, the goal of this study was to develop and validate a NLP model to identify professionalism lapses described in narrative evaluations of faculty completed by medical students, residents, and fellows. We hypothesized that a NLP approach could be used to reliably identify comments containing faculty professionalism lapses that are currently only obtainable with manual review.

Methods

Using a single-center retrospective cohort, we developed a text-based classification model to identify professionalism lapses in the narrative evaluations of faculty written by medical trainees (i.e., students, residents, and fellows).

Setting and Participants

We analyzed narrative comments from medical trainees' evaluations of clinical faculty at the University of Pennsylvania from July 2017 to December 2019.

The University of Pennsylvania evaluation system used in this analysis was developed in 2008 (McOwen et al., 2009). In its current version, the faculty evaluation included a cumulative numeric rating of global teaching effectiveness on a 5-point scale and narrative free-text comments. In 2017, in conjunction with the University of Pennsylvania's *Professionalism in Medicine* campaign, evaluations included an additional item to identify professionalism concern (Yes/No, with additional mandatory free text comments if "Yes").

Medical trainees completed evaluations at the end of each clinical rotation (both inpatient and outpatient), which occurred after variable exposure time to the faculty across different clinical departments.

The institutional review board (IRB) at the University of Pennsylvania determined the study to be exempt (#849815).

Data Collection and Text Processing

We used several NLP approaches in our model development. A total of 15,432 unique evaluations were included for analysis (which was 31% of all submitted clinical evaluations from July 2017-December 2019 [n = 48,384]). 12,444 evaluations were used for model development (using evaluations from 2017 and 2018), and 2988 for model validation (using evaluations from January 2019 to December 2019).

We constructed the training dataset using a random sample (n = 12,444, 25% of evaluations during the study period) of de-identified narrative evaluations and free-text comments associated with the professionalism item. The sample size was determined using a progressive sampling approach (Figueroa et al., 2012) in which the sample size increases until the model performance no longer improves with additional data.

We applied several different approaches to transform the free-text comments into a numeric format amenable to the development of classification models. First, we created an initial dictionary of terms and phrases indicative of professionalism lapses through an open review process (Schwartz et al., 2013) of 100 free-text comments that had previously been identified as having a professionalism concern. This keyword dictionary was iteratively amended based on review of model errors (both false positive and false negatives) in the training data, including the inclusion of negation terms in the dictionary (i.e., "not," "never, "none"). Supplemental Table 1 has a complete list of terms used for the keyword-based model. All terms and phrases were identified through group consensus.

In addition to the keyword dictionary, we used several additional techniques to transform text into formats amenable

to computational modeling (details of the complete approach are available in Supplemental Digital Appendix 1). First, we used term-frequency inverse document frequency weighted n-grams. This technique converts text into single words and phrases, and then quantifies the contextual importance of each phrase within the text by assigning each phrase a weight. Second, we used word embeddings. Word embeddings analyze semantic relationships across words and phrases and creates vectors to represent their meaning (such that words that are more similar in meaning have more similar vectors). Third, we used sentiment scoring which quantifies affective and subjective information in text, ranging on a continuous scale from −1 to 1, representing positive and negative attitudes, respectively.

Outcome Selection: Identification of Professionalism Lapses in Narrative Evaluations

To identify comments with a professionalism concern for use as the outcome (true positives) in our models, subject matter experts manually reviewed a subset of comments (JKH, CBC, JRK). The authors first formalized criteria for determining the presence or absence of unprofessional behavior in the clinical learning environment, which was informed by prior research (Ginsburg et al., 2002; Karnieli-Miller et al., 2010). The coders then manually reviewed narrative evaluations and coded for the presence or absence of a reference to a professionalism lapse.

To ensure reliability between coders, the coders initially coded 100 comments jointly and discussed discrepancies to reach consensus. The coding team then independently coded an additional 100 comments, and subsequently reviewed these 100 comments together.

After initial coding to ensure consensus, the remainder of the comments (n = 15,332) were reviewed by at least one of the coding team (with additional triple coding at set intervals to ensure ongoing reliability). Specifically, for coding the comments from the 2017-2018 academic year, 10% of the free-text comments were coded by the complete coding team (triple-coded). For comments from July 2018 to December 2019, 5% of the free-text comments were triple-coded. A total of 1073 comments were triple coded in the dataset. Iterative reliability checks using Cohen's kappa estimated >80% agreement between coders.

Model Training, Selection, and Assessment

To develop the classification models, we first fit the model on a subset of comments (to learn patterns in a portion of the dataset, "model training") and left the remaining data to examine how well it learned those patterns ("model testing"). We used the random sample described above from comments from 2017-2018 for model training and development, followed by 2988 randomly sampled comments from the 2019 academic year for model testing and temporal validation.

Using professionalism lapses identified on manual review as our outcome, we trained three types of models that are often employed in NLP-based categorization predictive modeling (Ng, 2004). These modeling approaches were selected to reduce overfitting, select the most important predictor variables, and eliminate insignificant covariates. Overfitting in a classification model would imply that the model had inappropriately learned random fluctuations in the data (noise) to make future predictions.

We tested for overfitting by comparing performance of the models in the training and testing sets. First, we used a penalized logistic regression model with L1 and L2 penalties. This approach penalizes the use of non-zero coefficients in from the regression model and "shrinks" their value toward zero if they do not contribute to improved performance, thereby creating less complex models that are also less likely to be overfit. Second, we used a random forest model which is a supervised machine learning algorithm that builds multiple decision trees using the most frequent results for classification. We then employed a neural network model which employs a set of algorithms designed to recognize patterns, including patterns of patterns, or representations of the data. Finally, we created and tested ensemble models, which incorporate each of these multiple regression models to improve the overall predictive performance. Additional details about the model tuning parameters and methods of analyzing model performance are outlined in the Supplemental Digital Appendix 1.

We assessed the model discrimination using the positive predictive value (PPV). PPV is the probability that a comment identified as a professionalism concern would be confirmed to have a true professionalism issue on manual review. (See Supplemental Table 2 for an example of application of PPV in this setting.) In this study, we aimed to create a NLP-based approach that identifies potential professionalism lapses in free-text comments. Therefore, for our purposes, we aimed for a high PPV to minimize the rate of false negatives, recognizing the tradeoff that the comments identified as potential professionalism concerns would require subsequent manual review. As an example, if approximately 1000 free-text comments were submitted per year, 240 would be flagged for professionalism concerns if there was a 20% prevalence of these issues. If the PPV was 67%, by the model and reviewed manually, 160 of these would be deemed "true positive" and lead to further action. (See Supplemental Table 2) Minimizing the false negatives in our model would still mean fewer comments needing manual review than our current approach of reviewing all completed evaluations.

We evaluated the PPV across the full range of classification thresholds for labeling comments as having professionalism concerns. A classification threshold is the cutoff at which the predicted probability is considered to be positive for the outcome of interest. In this case, when the predicted probability for a given sample of text was above that threshold, the model would predict the presence of a professionalism lapse. The optimal classification threshold in NLP-based predictive

modeling is determined using Precision-Recall curves, which assess the relationship between the true positive rate of professionalism lapses (recall) and the positive predictive value (precision) for a predictive model at different probability thresholds.

In addition to the PPV, we secondarily assessed the model with a calibration plot and the integrated calibration index (ICI), a score used to calculate the agreement between the observed and predicted outcome. We also calculated a scaled Brier score (Steyerberg et al., 2010), which provides a score for relative performance of the model predictions. A scaled Brier score quantifies the squared differences between actual outcomes and predictions, adjusting for the event rate (of professionalism lapses) in the population. This scaled Brier score does not have a target threshold. Instead, higher values indicate better performance relative to the base model (ranges from $-\infty$ to 1).

We completed all statistical analyses using the R language for statistical computing (version 4.0) and Python (version 3.8).

Results

Dataset/Demographic Characteristics

Faculty members represented all clinical departments within the School of Medicine. The majority of evaluations (n = 11,684, 76%) were submitted by graduate medical trainees (either in residency or fellowship training), and the remainder were submitted by medical students on clerkship, elective, and sub-internship rotations (Table 1).

Qualitative Identification of Outcome (Professionalism Concerns)

In total, the manual review of evaluations identified 551 comments with professionalism concerns across the full dataset (481 in 2017 and 2018 training set, and 70 in the remainder of the dataset). For outcome identification, the interrater reliability between the coding teams ranged from κ 0.71 to 1.0 (mean 0.87), indicating almost perfect agreement (Landis & Koch, 1977; Viera & Garrett, 2005).

Model Performance in the Validation Set

The performance characteristics of the top performing models (using the PPV, the scaled Brier score, and ICI) are outlined in Table 2.

The model using only the specific words (outlined in Supplemental Table 1) identified by the research team had moderate overall performance (scaled Brier score 0.14, 95% CI 0.06–0.22). This resulted in a PPV of 44% (CI 29%-58%), meaning 44% of cases identified by NLP-methods represented true professionalism lapses upon manual review. The model using sentiment analysis in combination with identified key words and negation terms (using the ensemble technique) performed best overall (scaled Brier score 0.36, 95% CI 0.26–0.48). The PPV of this model was 49% (CI 38%-59%), meaning that 49% of evaluations flagged using this model represented true professionalism lapses (based on manual review).

Figure 1 outlines the calibration plots and the precision-recall curves for the top performing predictive models.

Discussion

Our data highlight how a NLP-based classification model can be used to identify unprofessional faculty behavior in narrative evaluations. To our knowledge this is the first study to use NLP techniques to identify professionalism lapses noted in narrative evaluations of faculty in medical education. Given the challenges of identifying professionalism lapses in the clinical learning environment, including under-reporting through traditional mechanisms (Chung et al., 2018; Ginsburg et al., 2012; Hodges et al., 2011; Lucey & Souba, 2010; Mazer et al., 2018; Tucker et al., 2016), an automated NLP-based approach could provide an alternative method using already available narrative evaluations. Importantly, a NLP approach may more readily identify lapses noted in narrative evaluations of faculty, a benefit not possible through traditional reporting measures or through current time-intensive manual review of comments.

Our NLP-classification model incorporating key words, negation terms and sentiment analysis identified professionalism lapses in evaluation comments with a PPV of almost 50%. In other words, half of the comments identified as potential professionalism lapses remained "true positives" after manual review. We intentionally designed our NLP-classification approach to be oversensitive so that manual review efforts can focus on high-yield comments while minimizing false negatives. Therefore, while the NLP mechanism would not eliminate the necessary manual review of evaluations, it would significantly reduce the time burden of the current review process by flagging only a subset of evaluations for review. For example, if approximately 10,000 free-text comments were submitted per year, 400 would be flagged for professionalism concerns by the model and reviewed manually, of which 200 would be deemed "true positive" and lead to further action.

Importantly, the time savings in manual review must account for faculty time required for creation and maintenance of an NLP-classification system. Initial development of our NLP-classification system required approximately 10 hours of faculty time to review comments to identify key words for the NLP dictionary, and additional developer time to code and test the NLP algorithm. To implement this process at another institution, several hours of faculty time would be required upfront to review and adapt our NLP key word dictionary based on review of local comments, and to test and verify the NLP algorithm on their local comment dataset. After the initial development period, the NLP-algorithm requires approximately 2 hours per year of iterative

Table 1. Demographics of Cohort Used in Model Development and Validation.

	Total Dataset	Cohort Used in Model Development (Testing Dataset)	Cohort Used in Model Validation (Training Dataset)
Total evaluations (n, %)	15432 (100%)	12444 (81%)	2988 (19%)
Professionalism lapses identified[a] (n, %)	551 (4%)	481 (4%)	70 (2%)
Evaluations by level of trainee			
Residents/Fellows	11684 (76%)	9531 (77%)	2153 (72%)
Medical students	3700 (24%)	2865 (23%)	835 (28%)
Missing	48 (0%)	48 (0%)	0 (0%)
Evaluation numeric rating (mean, std dev)	3.64 (+/− 0.67)	3.64 (+/− 0.68)[b]	3.63 (+/− 0.68)
Total faculty[c] (n, %)	2470 (100%)	2211 (100%)	1432 (100%)
Faculty gender			
Female	1023 (41%)	897 (41%)	607 (42%)
Male	1304 (53%)	1194 (54%)	772 (54%)
Undisclosed/Missing	143 (6%)	120 (5%)	53 (4%)
Faculty race			
Asian	443 (18%)	389 (18%)	272 (19%)
Black	90 (4%)	77 (3%)	52 (4%)
Hispanic	62 (3%)	54 (2%)	37 (3%)
White	1662 (67%)	1519 (69%)	969 (68%)
Multiple	40 (2%)	33 (1%)	21 (1%)
Undisclosed/Missing	173(7%)	139 (6%)	81 (6%)
Faculty department			
Anesthesiology & critical care	212 (9%)	191 (9%)	134 (9%)
Dermatology	57 (2%)	49 (2%)	42 (3%)
Emergency medicine	81 (3%)	69 (3%)	64 (4%)
Family medicine	39 (2%)	34 (2%)	22 (2%)
Genetics	1 (0%)	2 (0%)	1 (0%)
Medicine	616 (25%)	542 (25%)	350 (24%)
Neurology	116 (5%)	105 (5%)	73 (5%)
Neurosurgery	20 (1%)	19 (1%)	8 (1%)
Obstetrics and gynecology	86 (3%)	74 (3%)	66 (5%)
Ophthalmology	33 (1%)	31 (1%)	17 (1%)
Orthopaedic surgery	49 (2%)	45 (2%)	30 (2%)
Otorhinolaryngology:	38 (2%)	34 (2%)	31 (2%)
Pathology	84 (3%)	76 (3%)	60 (4%)
Pediatrics	463 (19%)	430 (19%)	208 (15%)
Physical med. & rehabilitation	26 (1%)	23 (1%)	15 (1%)
Psychiatry	121 (5%)	98 (4%)	55 (4%)
Radiation oncology	36 (1%)	35 (2%)	13 (1%)
Radiology	160 (6%)	146 (7%)	118 (8%)
Surgery	154 (6%)	144 (7%)	104 (7%)
Non-clinical department[d]	4 (0%)	1 (0%)	3 (0%)
Undisclosed/Missing	74 (3%)	63 (3%)	18 (1%)

Notes. [a]Professionalism lapses identified by qualitative review.
[b]Numeric rating data on cohort used for model development based on 12,396 comments (missing rating data on 48 evaluations)
[c]Individual faculty members were represented in both the model development and testing cohort, for a total number of unique faculty in the overall cohort 2470.
[d]Clinical Departments include Biostatistics and Epidemiology, Cell and Developmental Biology, and Medical Ethics & Health Policy.

review to adjust the key word dictionary to account for secular changes in language use. Overall, for institutions with large evaluation databases, the time savings with implementation of an NLP-classification system is favorable compared to manual review, which require 1 hour of faculty time for review of 500-700 comments, plus reviewer training.

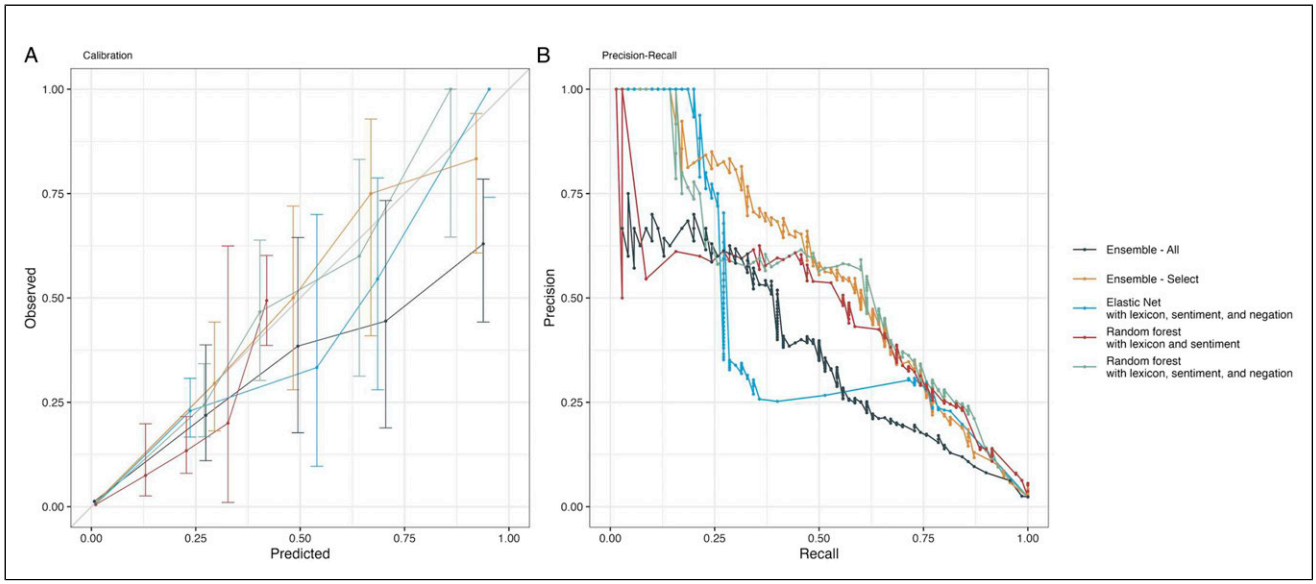
Even with creating a model designed to be oversensitive (and therefore intentionally identifying more comments as positive

Table 2. Model Performance Characteristics Across Top Performing Natural Language Processing-enhanced Models for Identification of Professionalism Lapses in SET.

	PPV[a] (95% CI)	Scaled Brier Score (95% CI)	ICI (95% CI)
Ensemble model			
All models	0.41 (0.30, 0.51)	0.18 (0.04, 0.34)	0.012 (0.007, 0.016)
Select[b] models	0.49 (0.38, 0.59)	0.36 (0.26, 0.48)	0.004 (−0.002, 0.005)
Elastic Net model			
Lexicon terms	0.39 (0.26, 0.53)	0.17 (0.08, 0.26)	0.007 (0.001, 0.010)
Lexicon with negation terms	0.37 (0.23, 0.50)	0.17 (0.08, 0.26)	0.007 (−0.0001, 0.010)
Lexicon with negation terms and sentiment	0.31 (0.23, 0.38)	0.34 (0.26, 0.43)	0.011 (0.006, 0.014)
Random forest model			
Lexicon terms	0.44 (0.29, 0.58)	0.14 (0.06 0.22)	0.011 (0.005, 0.015)
Lexicon and sentiment	0.44 (0.30, 0.58)	0.28 (0.21, 0.37)	0.011 (0.006, 0.015)
Lexicon and negation terms	0.44 (0.30, 0.58)	0.14 (0.07, 0.22)	0.011 (0.006, 0.014)
Lexicon, sentiment, and negation terms	0.40 (0.31, 0.48)	0.34 (0.26, 0.43)	0.008 (0.002, 0.010)
Term-frequency inverse document frequency (Tf-idf) Weighted N-grams	0.39 (0.26, 0.51)	0.11 (0.06, 0.14)	0.011 (0.006, 0.014)

Notes. [a]Positive predicted value.
[b]Models included in this ensemble model were the following: Elastic Net Model with lexicon terms and negation terms, Elastic Net Model with the vector and PCA approach, Elastic Net Model with sentiment score, lexicon terms, and negation terms, as well as the Random Forest Model with lexicon terms and negation terms, the Random Forest Model with the vector and PCA approach, and the Random Forest Model with sentiment score, lexicon terms, and negation terms.

Figure 1. 1(a) and (b) show the best performing Natural Language Processing models for identification of professionalism lapses in Narrative Evaluations. Figure A shows the calibration plot (comparison of the observed versus predicted probabilities of professionalism lapses) for top performing models. Figure B shows the precision-recall curves for top performing models, summarizing the relationship between the true positive rate of professionalism lapses (recall) and the positive predictive value (precision) for a predictive model across all probability thresholds.



using the NLP-algorithm and reducing the PPV), we suspect continued adjustments to the key word inclusion would be needed to improve accurate identification for future iterations. Our algorithm was designed to identify various combinations of language, including negative sentiment as well as specific key words (such as "disinterested," "avoids"). These word patterns could mistakenly identify evaluations that are instead negative teaching evaluations (but not necessarily professionalism lapses). We acknowledge the grey line between mistreatment and suboptimal learning environments and the subjective nature of the threshold for labeling behavior as unprofessional (Birden et al., 2014). While we attempted to address this issue by designing an

oversensitive model and allowing subsequent manual review to distinguish true professionalism concerns from suboptimal teaching, the process is still inherently subjective. Future directions for research include application of NLP-based classification models to recognize subtle language patterns or trends in language within individual faculty evaluation portfolios to detect early signs of unprofessional behavior, that may not have been detectable by manual review.

Of note, our manual review revealed a low overall prevalence of professionalism concerns (2–4% of evaluations), which highlights the importance of alternative reporting systems in identifying unprofessional faculty behavior and/or trainee mistreatment. However, given the significant barriers to reporting through formal mechanisms (Chung et al., 2018), review of these already available narrative comments (even with low incidence) provide a supplemental source for identifying unprofessional faculty behavior. With low overall prevalence of comments containing references to unprofessional behavior, the NLP-based screening mechanism serves to concentrate time and resources to high-yield manual review.

One surprising study result was that the addition of more advanced NLP approaches (including term-frequency inverse document frequency weighted n-grams, which aims to quantify the contextual importance of each phrase within text, and vector development, which analyzes relationships across words and sentences) did not provide sufficient additive benefit to the model. This suggests that future work should focus on optimizing the key word dictionary and algorithm rather than incorporating new NLP models.

Study strengths include the inclusion of narrative evaluation data derived from all clinical departments within the University of Pennsylvania, with learners from across the education continuum (i.e., students, residents, and fellows). The use of manual qualitative coding to determine the outcome in model creation provided a critical gold standard for model development. Generalizability is limited as our dataset included faculty within a single academic institution. Furthermore, the language used to describe professionalism issues is heterogeneous (with considerable variation across and within individuals), and the manifestations of unprofessional behavior are diverse. Therefore, certain unprofessional behaviors may not be captured if they were not represented in the current dataset. Additionally, the language used to describe professionalism concerns may also change over time due to cultural shifts. These limitations highlight the need for regular iterative review and ongoing dictionary and model refinement to ensure the model continues to capture concerning behaviors.

In conclusion, this study identified a scalable, automated strategy for screening narrative evaluations of faculty to identify professionalism lapses. Overall, a NLP-based classification model combining key words, negation terms, and sentiment analysis flagged comments for professionalism concerns with a PPV of 49%. This novel, time- and resource-efficient approach for identifying potential professionalism lapses may ultimately be employed as one of a suite of tools to address faculty professionalism issues and potentially improve the clinical learning environment.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: University of Pennsylvania Department of General Internal Medicine Sam Martin Institutional Award.

ORCID iD

Janae K. Heath https://orcid.org/0000-0002-0533-3088

Supplemental Material

Supplemental material for this article is available online.

References

AAMC Graduation Questionnaire (GQ). (2021). Association of American Medical Colleges. https://www.aamc.org/media/55736/download

Birden, H., Glass, N., Wilson, I., Harrison, M., Usherwood, T., & Nass, D. (2014). Defining professionalism in medical education: A systematic review. *Medical Teacher*, *36*(1), 47–61. https://doi.org/10.3109/0142159X.2014.850154

Chary, M., Parikh, S., Manini, A. F., Boyer, E. W., & Radeos, M. (2019). A review of Natural Language Processing in medical education. *The Western Journal of Emergency Medicine*, *20*(1), 78–86. https://doi.org/10.5811/westjem.2018.11.39725

Chung, M. P., Thang, C. K., Vermillion, M., Fried, J. M., & Uijtdehaage, S. (2018). Exploring medical students' barriers to reporting mistreatment during clerkships: A qualitative study. *Medical Education Online*, *23*(1), 1478170. https://doi.org/10.1080/10872981.2018.1478170

Cook, A. F., Arora, V. M., Rasinski, K. A., Curlin, F. A., & Yoon, J. D. (2014). The prevalence of medical student mistreatment and its association with burnout. *Academic Medicine: Journal of the Association of American Medical Colleges*, *89*(5), 749–754. https://doi.org/10.1097/ACM.0000000000000204

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. https://doi.org/10.1186/1472-6947-12-8

French, J. C., Dannefer, E. F., & Colbert, C. Y. (2014). A systematic approach toward building a fully operational clinical competency committee. *Journal of Surgical Education*, *71*(6), Aricle e22–e27. https://doi.org/10.1016/j.jsurg.2014.04.005

Gan, R., & Snell, L. (2014). When the learning environment is suboptimal: Exploring medical students' perceptions of "mistreatment. *Academic Medicine: Journal of the Association of*

American Medical Colleges, *89*(4), 608–617. https://doi.org/10.1097/ACM.0000000000000172

Gierl, M. J., Latifi, S., Lai, H., Boulais, A.-P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, *48*(10), 950–962. https://doi.org/10.1111/medu.12517

Ginsburg, S., Bernabeo, E., Ross, K. M., & Holmboe, E. S. (2012). "It depends": Results of a qualitative study investigating how practicing internists approach professional dilemmas. *Academic Medicine: Journal of the Association of American Medical Colleges*, *87*(12), 1685–1693. https://doi.org/10.1097/ACM.0b013e3182736dfc

Ginsburg, S., Regehr, G., Stern, D., & Lingard, L. (2002). The anatomy of the professional lapse: Bridging the gap between traditional frameworks and students' perceptions. *Academic Medicine: Journal of the Association of American Medical Colleges*, *77*(6), 516–522. https://doi.org/10.1097/00001888-200206000-00007

Heath, J. K., & Dine, C. J. (2019). ACGME milestones within subspecialty training programs: One institution's experience. *Journal of Graduate Medical Education*, *11*(1), 53–59. https://doi.org/10.4300/JGME-D-18-00308.1

Hodges, B. D., Ginsburg, S., Cruess, R., Cruess, S., Delport, R., Hafferty, F., Ho, M.-J., Holmboe, E., Holtman, M., Ohbu, S., Rees, C., Ten Cate, O., Tsugawa, Y., Van Mook, W., Wass, V., Wilkinson, T., & Wade, W. (2011). Assessment of professionalism: Recommendations from the ottawa 2010 conference. *Medical Teacher*, *33*(5), 354–363. https://doi.org/10.3109/0142159X.2011.577300

Hu, Y.-Y., Ellis, R. J., Hewitt, D. B., Yang, A. D., Cheung, E. O., Moskowitz, J. T., Potts, J. R., Buyske, J., Hoyt, D. B., Nasca, T. J., & Bilimoria, K. Y. (2019). Discrimination, abuse, harassment, and burnout in surgical residency training. *The New England Journal of Medicine*, *381*(18), 1741–1752. https://doi.org/10.1056/NEJMsa1903759

Karnieli-Miller, O., Vu, T. R., Holtman, M. C., Clyman, S. G., & Inui, T. S. (2010). Medical students' professionalism narratives: A window on the informal and hidden curriculum. *Academic Medicine: Journal of the Association of American Medical Colleges*, *85*(1), 124–133. https://doi.org/10.1097/ACM.0b013e3181c42896

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lucey, C., & Souba, W. (2010). Perspective: The problem with the problem of professionalism. *Academic Medicine: Journal of the Association of American Medical Colleges*, *85*(6), 1018–1024. https://doi.org/10.1097/ACM.0b013e3181dbe51f

Mazer, L. M., Bereknyei Merrell, S., Hasty, B. N., Stave, C., & Lau, J. N. (2018). Assessment of programs aimed to decrease or prevent mistreatment of medical trainees. *JAMA Network Open*, *1*(3), Article e180870. https://doi.org/10.1001/jamanetworkopen.2018.0870

McOwen, K. S., Bellini, L. M., Morrison, G., & Shea, J. A. (2009). The development and implementation of a health-system-wide evaluation system for education activities: Build it and they will come. *Academic Medicine: Journal of the Association of American Medical Colleges*, *84*(10), 1352–1359. https://doi.org/10.1097/ACM.0b013e3181b6c996

Ng, A. Y. (2004). Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In Twenty-first International Conference on Machine Learning - ICML '04, 2004. https://doi.org/10.1145/1015330.1015435

Oser, T. K., Haidet, P., Lewis, P. R., Mauger, D. T., Gingrich, D. L., & Leong, S. L. (2014). Frequency and negative impact of medical student mistreatment based on specialty choice: A longitudinal study. *Academic Medicine: Journal of the Association of American Medical Colleges*, *89*(5), 755–761. https://doi.org/10.1097/ACM.0000000000000207

Richman, J. A., Flaherty, J. A., Rospenda, K. M., & Christensen, M. L. (1992). Mental health consequences and correlates of reported medical student abuse. *JAMA: The Journal of the American Medical Association*, *267*(5), 692–694. https://doi.org/10.1001/jama.1992.03480050096032

Ross, P. T., Abdoler, E., Flygt, L., Mangrulkar, R. S., & Santen, S. A. (2018). Using a modified A3 lean framework to identify ways to increase students' reporting of mistreatment behaviors. *Academic Medicine: Journal of the Association of American Medical Colleges*, *93*(4), 606–611. https://doi.org/10.1097/ACM.0000000000002033

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Tremblay, G., Carmichael, P.-H., Maziade, J., & Grégoire, M. (2019). Detection of residents with progress issues using a keyword–specific algorithm. *Journal of Graduate Medical Education*, *11*(6), 656–662. https://doi.org/10.4300/JGME-D-19-00386.1

Tucker, C. R., Choby, B. A., Moore, A., Parker, R. S., Zambetti, B. R., Naids, S., Scott, J., Loome, J., & Gaffney, S. (2016). Speaking up: Using OSTEs to understand how medical students address professionalism lapses. *Medical Education Online*, *21*(1), 32610. https://doi.org/10.3402/meo.v21.32610

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Wilkinson, T. J., Gill, D. J., Fitzjohn, J., Palmer, C. L., & Mulder, R. T. (2006). The impact on students of adverse experiences during medical school. *Medical Teacher*, *28*(2), 129–135. https://doi.org/10.1080/01421590600607195

Wilkinson, T. J., Wade, W. B., & Knock, L. D. (2009). A blueprint to assess professionalism: Results of a systematic review. *Academic Medicine: Journal of the Association of American Medical Colleges*, *84*(5), 551–558. https://doi.org/10.1097/ACM.0b013e31819fbaa2