# Automated Assessment of Medical Training Evaluation Text

**Rui Zhang, MS[1], Serguei Pakhomov, PhD[1, 2], Sophia Gladding, PhD[3], Michael Aylward, MD[4], Emily Borman-Shoap, MD[3], and Genevieve B. Melton, MD, MA[1,5]**
**[1]Institute for Health Informatics; [2]College of Pharmacy; and Departments of [3]Pediatrics, [4]Medicine, and [5]Surgery; University of Minnesota, Minneapolis, MN**

**Abstract**

*Medical post-graduate residency training and medical student training increasingly utilize electronic systems to evaluate trainee performance based on defined training competencies with quantitative and qualitative data, the later of which typically consists of text comments. Medical education is concomitantly becoming a growing area of clinical research. While electronic systems have proliferated in number, little work has been done to help manage and analyze qualitative data from these evaluations. We explored the use of text-mining techniques to assist medical education researchers in sentiment analysis and topic analysis of residency evaluations with a sample of 812 evaluation statements. While comments were predominantly positive, sentiment analysis improved the ability to discriminate statements with 93% accuracy. Similar to other domains, Latent Dirichlet Analysis and Information Gain revealed groups of core subjects and appear to be useful for identifying topics from this data.*

## Introduction

A growing area of clinical research is the area of medical education, which specifically focuses on ways to advance the pedagogy and methods to assess the knowledge, skills, and attitudes of medical students, residents, and other clinical trainees[1-3]. Assessment of trainee performance in medical education is competency-based; however, the specific metrics and predictors of success of practicing clinicians remain poorly defined. For physician trainees in particular, the relationship between a given resident's total body of evaluations and their educational and career outcomes remains elusive. Training regulations, including the specific training competencies, are governed by the Accreditation Council for Graduate Medical Education (ACGME)[4]. The decision by the ACGME to move all graduate medical education assessment to a competency-based system has resulted in significant changes to the approach to resident evaluation.

In order to comply with the ACGME competency framework, residency programs have developed evaluation tools that are often delivered and compiled electronically. The delivery and review of these more detailed evaluations places additional administrative burdens on residency administrative staff and faculty at a time when there are added pressures from changes in work hour regulations and curricular requirements. In addition, changes in the federal government funding model have resulted in more limited funding resources for programs[5-7]. The individual components of these evaluations are most commonly at the discretion of the residency programs, and there is little data to direct the creation of these evaluations. Electronic evaluation systems allow for both quantitative ratings and qualitative (text) ratings of performance and are typically organized according to each of the six ACGME core competencies which provide a framework[8]. While these systems allow for quantitative data to be easily compiled for summarization of the individual trainee compared to other residents, some studies have noted that this data can be predominantly positively biased and rater-specific due to differences in how ratings are done by each individual faculty instructor[9]. Furthermore, some ACGME competencies, as depicted in Table 1, can be difficult to measure by individual evaluators and for trainees to learn[10-12]. While some methodologies have been developed to better assess these competencies in residency training, these techniques can be highly time intensive for evaluators and not easily scalable for analysis within large training programs[13]. The qualitative text currently collected is part of the written comments on evaluation forms and is generally not synthesized in any systematic way. Methods to review and synthesize the information of these comments as an additional measure of trainee evaluation remain relatively unexplored, keeping this data relatively "locked" in text format and not available as a comparative tool.

We sought to explore text mining techniques applied to a number of different domains previously[14] to assess the feasibility and value of an automated approach for synthesizing evaluation comments of residency trainees to aid research in medical education and the ability of residency programs to make better assessments of trainees. While these techniques were developed within general computational linguistics, little has been reportedly in their use within the medical domain particularly for medical education evaluation texts. The goal of this study was to assess the sentiment or opinion expressed and topic(s) of evaluation texts with automatic text mining techniques.

**Table 1.** ACGME Core Competencies[4].

| Competency | Definition (exact wording of ACGME) |
|---|---|
| Patient Care | Residents must be able to provide patient care that is compassionate, appropriate, and effective for the treatment of health problems and the promotion of health. |
| Medical Knowledge | Residents must demonstrate knowledge of established and evolving biomedical, clinical, epidemiological and social-behavioral sciences, as well as the application of this knowledge to patient care. |
| Practice-Based Learning and Improvement | Residents must demonstrate the ability to investigate and evaluate their care of patients, to appraise and assimilate scientific evidence, and to continuously improve patient care based on constant self-evaluation and life-long learning. Residents are expected to develop skills and habits to be able to meet the following goals:<br>• identify strengths, deficiencies, and limits in one's knowledge and expertise;<br>• set learning and improvement goals;<br>• identify and perform appropriate learning activities;<br>• systematically analyze practice using quality improvement methods, and implement changes with the goal of practice improvement;<br>• incorporate formative evaluation feedback into daily practice;<br>• locate, appraise, and assimilate evidence from scientific studies related to their patients' health problems;<br>• use information technology to optimize learning; and,<br>• participate in the education of patients, families, students, residents and other health professionals. |
| Systems-Based Practice | Residents must demonstrate an awareness of and responsiveness to the larger context and system of health care, as well as the ability to call effectively on other resources in the system to provide optimal health care. Residents are expected to:<br>• work effectively in various health care delivery settings and systems relevant to their clinical specialty;<br>• coordinate patient care within the health care system relevant to their clinical specialty;<br>• incorporate considerations of cost awareness and risk-benefit analysis in patient and/or population-based care as appropriate;<br>• advocate for quality patient care and optimal patient care systems;<br>• work in interprofessional teams to enhance patient safety and improve patient care quality; and,<br>• participate in identifying system errors and implementing potential systems solutions. |
| Professionalism | Residents must demonstrate a commitment to carrying out professional responsibilities and an adherence to ethical principles. Residents are expected to demonstrate:<br>• compassion, integrity, and respect for others;<br>• responsiveness to patient needs that supersedes self-interest;<br>• respect for patient privacy and autonomy;<br>• accountability to patients, society and the profession; and,<br>• sensitivity and responsiveness to a diverse patient population, including but not limited to diversity in gender, age, culture, race, religion, disabilities, and sexual orientation. |
| Interpersonal Skills and Communication | Residents must demonstrate interpersonal and communication skills that result in the effective exchange of information and collaboration with patients, their families, and health professionals. Residents are expected to:<br>• communicate effectively with patients, families, and the public, as appropriate, across a broad range of socioeconomic and cultural backgrounds;<br>• communicate effectively with physicians, other health professionals, and health related agencies;<br>• work effectively as a member or leader of a health care team or other professional group;<br>• act in a consultative role to other physicians and health professionals; and,<br>• maintain comprehensive, timely, and legible medical records, if applicable |

## Background

### *Sentiment analysis*

Sentiment analysis or opinion mining is a natural language processing technique that determines the attitude of the evaluation or the polarity of a comment or text. Automatic sentiment analysis has been used in several different domains such as banking, travel, movie, and automobile reviews. These techniques are generally classified into two categories: lexical-based techniques and machine learning-based techniques.

Lexical-based techniques include a bag-of-words approach without consideration of relations between words has been widely used. Sentiments of text have been calculated with an aggregation function from semantic orientation of all words in the text. Minimal path length (MPL) has been used to measure the distance of words in WordNet to determine the semantic orientation of a word[15]. Some lexical-based methods used only adjectives as the indicator of semantic orientation of text[16]. The combinations of lexical word combinations such as adjective-nouns (AN)[17], adverb-adjective combination (AAC)[18] and adjective-verb-adverb (AVA)[19] have also been used as the indicators to evaluate the sentiment. In a more complex technique, which considers relationships between words, Mulder combined lexical-based techniques with a simple grammar to formalize affective words with their objects[20].

With machine learning-based techniques, standard supervised methods used for sentiment analysis including Support Vector Machines (SVM), Naïve Bayes (NB) and Maximum Entropy (ME)[21, 22]. Features have included *N*-grams, lexical normalization, part-of-speech (POS), negation[23], and opinion words[24]. Machine learning-based techniques have shown to be more effective than lexical-based techniques. As with other applications of supervised learning, these features are implemented with a training set of data to build the model and then tested on a validation set to assess performance.

### *Topic analysis*

Topic analysis includes topic identification and text segmentation, which extracts thematic structure from a text[25]. Topic analysis helps to identify the subtopics of a text and can detect opinion changes over subtopics in a piece of text. An early topic and formative model, Latent Semantic Analysis (LSA), analyzes the relationship between documents and terms based on the position of the words in the texts[26]. Within current literature, Latent Dirichlet Allocation (LDA) is one of the more common topic model techniques[27]. LDA is a statistical model, which assumes that a text may exhibit multiple topics, which are probability distributions over words. LDA is a generative graphic model, which can discover the underlying topic structures of texts. The document generation process is shown as shown in Figure 1. Each document picks a multinomial distribution $\theta$ from a Dirichlet distribution with a parameter $\alpha$. Each word $w$ in a document $i$ choose a topic $z$ from $\theta$. Each topic $z$ picks a multinomial distribution $\phi$ from another parameter $\beta$ of Dirichlet distribution. Finally, the model picks word $w$ from $\phi$.

Feature selection is a technique widely used to find the top affect features on topic classification based on feature ranking metrics. Keywords of each topic obtained from topic-term matrix of LDA can provide good feature candidates, since the LDA-based keywords are important for uncovering latent topics. Another conventional feature selection metric is information gain (IG), which tests entropy changes of the system (e.g., topic classification) with and without a feature (e.g., word); thus, IG can help select top important features to the system.

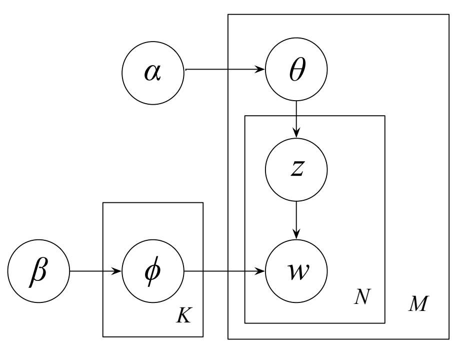


**Figure 1.** Graphic model of LDA. $\alpha$, parameter of the Dirichlet distribution on the per-document topic distributions; $\beta$, parameter of the Dirichlet distribution on the per-topic word distribution; $\phi_k$ word distribution of a topic $k$; $\theta_i$, topic distribution of a document $i$; $z_{ij}$, topic of $j$th word in document $i$; $w_{ij}$, the $j$th word in document $i$; $M$, the total number of documents; $N$, a given document; $K$, the number of latent topics.

## Methods

The objectives of this study were to explore text mining techniques with medical training text in order to: a) analyze the sentiment or opinion (in three categories: extremely positive, positive and non-positive) of each evaluation statement; b) assign ACGME core competencies (Table 1) to each evaluation statement. The overall approach (Figure 2) involved four phrases: (1) collect medical education evaluation statements; (2) apply machine learning algorithms (e.g., NB, SVM) with selected features to analyze sentiment of each statement; 3) utilize LDA, IG, POS features with SVM to classify each statement to one topic; (4) build up manually annotated reference standard based on which the methods were evaluated.

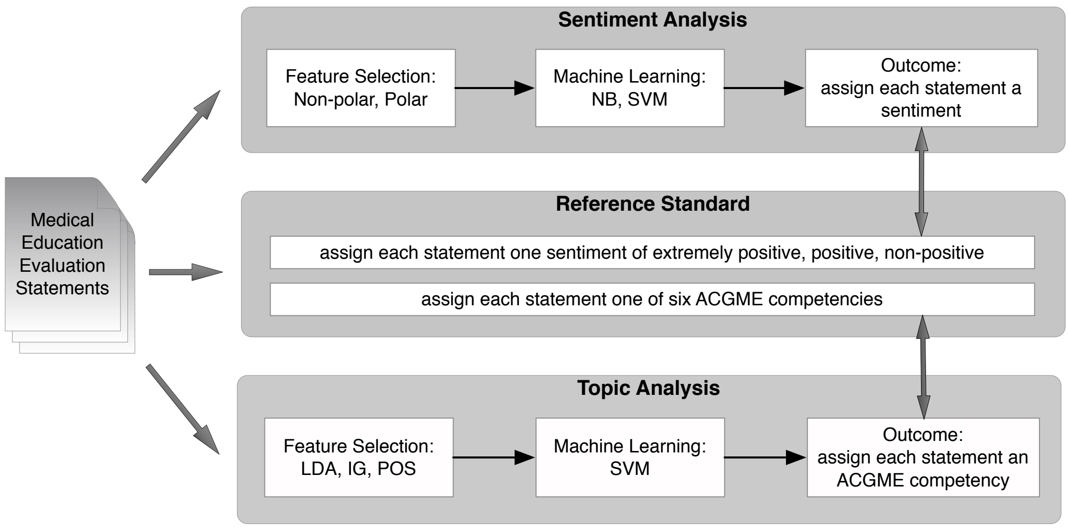


**Figure 2.** Overview of approach. NB, naïve bayes; SVM, support vector machine; LDA, latent Dirichlet allocation; IG, information gain; POS, part-of-speech.

### *Data collection*

Electronic evaluations of a subset of residents completed by faculty in the Pediatrics and Medicine-Pediatrics residencies at the University of Minnesota program were utilized from the year 2008 for this pilot study. Evaluations were created and distributed using the New Innovations RMS Evaluations application[28]. The overall set of evaluations from this system includes 360-degree evaluations of residents from faculty, peers, and staff members, as well as residents evaluating faculty and others. For the purposes of this pilot study, a subset of evaluation comments (812 statements) from faculty evaluating residents was used in order to maintain uniformity of the formative dataset. University of Minnesota institutional review board approval was obtained and informed consent waived in this minimal risk study.

Each evaluation includes quantitative performance outcomes based on the ACGME competencies including medical knowledge (MK), patient care (PC), practice-based learning and improvement (PBLI), interpersonal communication skills (ISC), professionalism (PRO) and system-based practice (SBP), along with a general evaluation of the trainee. Similarly, faculty comments of residents for each category and overall were typed into the Web-based software application in free-text format. For the purposes of this pilot, a subset of evaluations of final year residents was utilized so as to maintain and hold-out the evaluations of current residents for a future study where correlation with performance outcomes will be made.

### *Sentiment analysis*

We utilized supervised machine learning algorithms, NB and SVM, using *non-polar* feature sets based on previous work including: unigram, unigram with lexical normalization, adjective (JJ), adverb (VB), noun (NN) and their combinations. Part-of-speech (POS) tags were determined by using Stanford parser[29]. For the purpose of this study, we also extended methods to the *polar* features by using Multi-Perspective Question Answering (MPQA) subjectivity lexicon, a list of 8221 words with their corresponding polarities and strengths (e.g., "excellent" is identified as a strong positive word)[30]. To implement each algorithm, each resident evaluation text statement was represented as the statement vector of features. For example, $(f_1, f_2, \ldots, f_m)$ is a set of features. Let $n_i$ be the presence of $f_i$ in statement $s_j$ (e.g., if $f_i$ is present in a statement $s_j$, $n_{ij} = 1$; otherwise, $n_{ij} = 0$), then the statement vector = ($n_1$, $n_2$ …, $n_m$). Algorithms were implemented using Weka and validated using 10-fold cross validation[31]. For the purposes of this initial analysis, sentiment was evaluated as 3-way (extremely positive (EP) versus positive (P) versus non-positive (NP)). The category NP consisted of all neutral (U), negative (N) and extremely negative (EN) statements as evaluated by raters.

***Topic analysis***

We chose three methods, LDA, IG and POS, for feature selection and then used supervised SVM machine learning method to classify statements to six competencies (i.e., PC, MK, PBLI, SBP, PRO, ISC). Only 699 statements were selected for topic analysis since 113 were not annotated in our expert standard to map to one of the competencies. We first removed stopwords[32] and used lexical variation generation (LVG)[33] to lexically normalized all word tokens.

LDA with Gibbs Sampling (iteration = 1500) was implemented in Stanford Topic Model Toolkit (TMT-0.4.0)[34]. Topic numbers from 50-500 were chosen. Keywords for each topic were obtained from the output files. We then used IG to rank those keywords for each topic and chose filtered top numbers of keywords (from 50-350) as features to implement classification.

We then used IG to reduce word features from all words in the whole training evaluation texts. We also used Stanford Parser to find POS tags for each sentences. We only utilized adjective (JJ) and noun (NN) as features as they have previously been shown to contain important information related to topics. We finally implemented SVM based on these features and evaluated the performance by reporting accuracy, precision, recall, and F-measure on each competency using 10-fold cross validation.

***Manually annotated reference standard***

Two hundred and thirty-three comments from resident evaluations were selected for this study and split into 812 individual statements (incomplete sentences) and sentences. Two native English speakers with knowledge of residency training and evaluation system were asked to rate each sentence into five categories (i.e., EP, P, U, N, EN), and to assign one competency (i.e., PC, MK, PBLI, SBP, PRO, ISC) for each sentence or statement. Inter-rater reliability of two annotators at a sentence level was determined by percent agreement[35] for a portion of the statements from the dataset (80 of 812 statements). With the topic assignment standard, consensus was reached for topics on sentences where a clear ACGME competency could not be clearly assigned. Performance was determined at a sentence/statement level.

## Results

A total of 812 sentences were available for this study, with 55.4% of the comments having an extremely positive sentiment and 32.9% a positive sentiment. Two experts completely agreed on rating sentiment (100%) for the overlap sample 80 statements (ten percent of the whole dataset). Most statements were categorized as "ISC" (n = 221), "PRO" (n = 148), and "PBLI" (n = 116).

***Sentiment analysis***

Evaluation statements were overwhelmingly positive (EP + P). Over half of the statements are EP, resulting in low precision and recall of baseline, as shown in Table 2. The adding features of adjective (JJ), verb (VB), adverb (RB) and noun (NN) helped to improve the performance. Subjective (polar) words (SW) were important for correctly assigning sentiment. The combined features of adjective (JJ) + adverb (RB) + verb (VB) + noun (NN) + subjective words (SW) with SVM achieved the best accuracy (93.7%) for 3-way sentiment classification.

**Table 2.** Performance of sentiment 3-way classification (EP versus P versus NP) on resident evaluation text.

| Features | Accuracy | | Precision | | Recall | | F-Measure | |
|---|---|---|---|---|---|---|---|---|
| | NB | SVM | NB | SVM | NB | SVM | NB | SVM |
| Baseline | 0.702 | | 0.307 | | 0.554 | | 0.359 | |
| JJ | 0.799 | 0.855 | 0.691 | 0.796 | 0.698 | 0.782 | 0.764 | 0.888 |
| SW | 0.833 | 0.894 | 0.754 | 0.865 | 0.750 | 0.841 | 0.728 | 0.838 |
| Unigram | 0.828 | 0.927 | 0.753 | 0.893 | 0.742 | 0.890 | 0.722 | 0.888 |
| JJ + RB +VB | 0.825 | 0.930 | 0.752 | 0.899 | 0.737 | 0.895 | 0.715 | 0.893 |
| JJ + RB +VB + NN | 0.825 | 0.933 | 0.751 | 0.905 | 0.737 | 0.900 | 0.712 | 0.898 |
| JJ + RB +VB + NN + SW | 0.845 | **0.937** | 0.775 | 0.909 | 0.768 | 0.905 | 0.751 | 0.902 |

*NB, naïve bayes; SVM, support vector machine; JJ, adjective; SW, subjective words; RB, adverb; VB, verb; NN, noun.

***Topic analysis***
In LDA, the performance (proportion of correctly predicted topics) of the classifier overall improved with increasing the numbers of topics (Figure 3(A)). Keywords with 400 topics appeared to present the best performance for almost all numbers of keywords over 100. In general, the peak of performance for each curve appears to be between 100 and 200 keywords. The feature set of 150 keywords with 400 topics performed the best overall in classification of all competencies. As shown in Figure 3(B), the performances of topic analysis with different numbers of keywords varied significantly with IG. The best performance was reached with approximately 350 keywords.

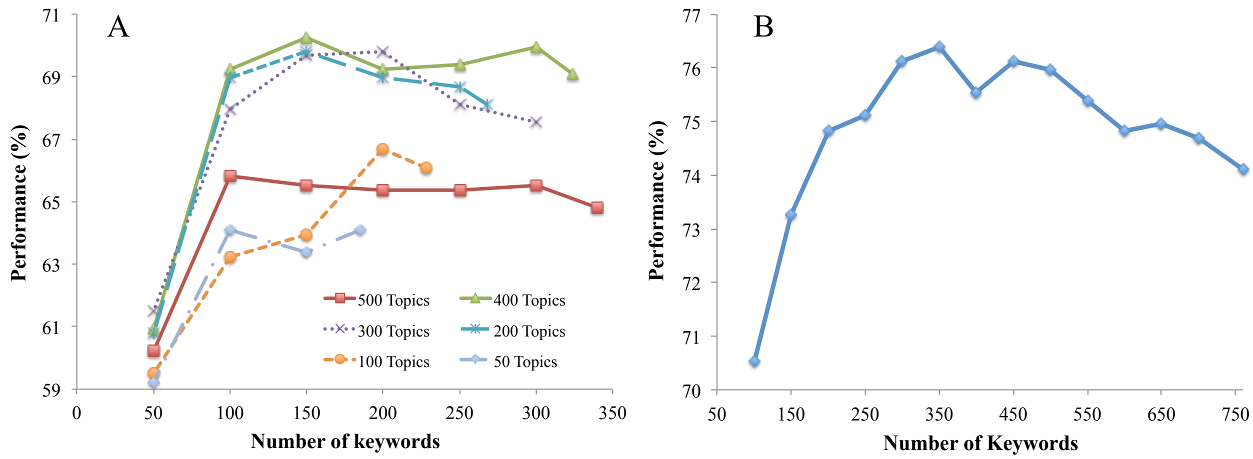


**Figure 3.** Performance of SVM models with A) various numbers of topics in LDA; and B) various numbers of keywords using IG.

Using the two most effective features (adjective (JJ) + noun (NN)), the best recall was 75.9% with IG (Table 3). Measures of accuracy, precision, recall and F-measure for each competency were also calculated with each method.

**Table 3.** Performance of topic analysis on resident evaluation text using SVM.

| Features | Competency | Precision | Recall | F-Measure |
|---|---|---|---|---|
| LDA (150 words) | PC | 0.429 | 0.286 | 0.343 |
| | MK | 0.784 | 0.690 | 0.734 |
| | PBLI | 0.871 | 0.638 | 0.736 |
| | SBP | 0.694 | 0.490 | 0.575 |
| | PRO | 0.800 | 0.649 | 0.716 |
| | ISC | 0.637 | 0.946 | 0.761 |
| | OVERALL | 0.717 | 0.702 | 0.692 |
| JJ + NN (524 words) | PC | 0.647 | 0.698 | 0.672 |
| | MK | 0.809 | 0.720 | 0.762 |
| | PBLI | 0.780 | 0.672 | 0.722 |
| | SBP | 0.622 | 0.451 | 0.523 |
| | PRO | 0.654 | 0.804 | 0.721 |
| | ISC | 0.798 | 0.805 | 0.802 |
| | OVERALL | 0.740 | 0.735 | 0.734 |
| IG (350 words) | PC | 0.639 | 0.730 | 0.681 |
| | MK | 0.806 | 0.750 | 0.777 |
| | PBLI | 0.750 | 0.672 | 0.709 |
| | SBP | 0.667 | 0.431 | 0.524 |
| | PRO | 0.721 | 0.838 | 0.775 |
| | ISC | 0.836 | 0.851 | 0.843 |
| | OVERALL | 0.763 | 0.763 | 0.759 |

* SVM, support vector machine; LDA, latent Dirichlet allocation; JJ, adjective; NN, noun; IG, information gain; PC, patient care; MK, medical knowledge; PBLI, practice-based learning and improvement; SBP, system-based practice; PRO, professionalism; ISC, interpersonal skills and communication; OVERALL, overall performance for all competencies. Precision=TruePositive/(TruePositive+NegativePositive); Recall=TruePositive/(TruePositive+TrueNegative); F-measure=2×Precision×Recall/(Precision + Recall).

Several subtopics for sentences of ISC (the competency contained the largest number of sentences) were also found and then manually verified. Several of the topics, keywords and example sentences are listed in Table 4.

**Table 4.** Subtopics for the competency of "Interpersonal Skills and Communications".

| Topic#<br>Assigned Name | Keywords | Examples |
|---|---|---|
| 1<br>Teamwork | Team<br>Member<br>Impact | Communicates effectively with all members of the health care team<br>NAME's positive energy can greatly impact a team<br>Lack of communication with fellows |
| 2<br>Role model | Role Model<br>Player | Outstanding role model for students<br>Excellent team player |
| 3<br>Interpersonal and communication skills | Skill<br>Interpersonal<br>Communication | Excellent communication skills<br>Excellent interpersonal skills<br>Great interpersonal communication skills |
| 4<br>Team leadership and supervision | Leader<br>Supervision | Excellent team leader<br>She tried to foster teamwork with her fellow residents<br>Excellent supervision of junior residents |
| 5<br>Personality | Personality<br>Humor | Very warm personality<br>Several also commented on her great personality<br>As usual, she was complimented for being fun to work with |

**Discussion**

This study explores text-mining techniques aimed to help medical educators automatically analyze medical education evaluation texts in two tasks: sentiment (opinion) analysis and topic analysis of statements. While there has been little focused work on qualitative evaluation data, it is an important and useful research area. In the increasingly important area of medical education research, greater regulation and resource limitations from an over-burdened residency training system, as well as a competency-based accreditation system, the assessment of resident trainee progress is increasingly critically important and complex for residency programs. Medical educators are faced with large amounts of unsynthesized qualitative text evaluation data that is often without context and which currently requires tedious review with unclear correlations to resident outcomes. Synthesis of this data is also made more difficult by the predominantly positive nature of comments and quantitative ratings, making it challenging to identify below average, average, or excellent trainees, the former two of which might benefit from early identification and focused interventions for remediation. Furthermore, the distribution of the data across several years of training, and different evaluators and settings makes identifying trends and context specific deficiencies (or aptitudes) difficult. With the use of a small dataset of over 800 evaluation statements, we demonstrate that traditional natural language processing sentiment analysis approaches and feature sets are promising techniques for analyzing text in the medical education domain. Using standard statistical-based models to assign statements to topic(s), the LDA technique and information gain identified core topics and themes within statements.

When using different machine learning techniques and features to identify sentiment of statements, we found that SVM outperformed NB with each feature set for 3-way classification. While our corpus was limited by the relative small number of negative and neutral statements to train our models, this is a practical task needed to evaluate the performance of residents. We believe that differentiating EP and P is more valuable in the resident evaluation tasks than in other domains (e.g., movie and automobile reviews) due to the predominately positive nature of medical evaluation text. In the medical residency training domain, one may be able to equalize the classes of EP, P and NP as excellent, average and below average.

One particular challenge of this task is to classify not only the polarity of the word (e.g., "good" (P) versus "bad" (N)), but also the strength of the word (e.g., "good" (P) versus "excellent" (EP)) or the combination of RB and JJ (e.g., "good" (P) versus "extremely good" (EP)). The negation words (e.g., "not", "cannot") can also completely change the sentiment of whole statement and sentence. Some verbs also have polarity, such as "lack" (N). Our results show that adding adverb (RB) and verb (VB) into the feature set achieved a better accuracy (improvement of 7.5%) compared to use of only adjective (JJ). Nouns can also have sentiment, such as "star" (EP) as in the sentence, "he is a star resident". However, adding noun (NN) to our feature set did not significantly improve the performance likely because only a small number of nouns with polarity are likely contained in such a small dataset. Subjective words with the corresponding strength can also help the classifier to differentiate EP versus P. While unigram may achieve good performance for some tasks, here it was not particularly effective since it included every word into the feature set and was computationally intensive. Future studies with larger datasets will be needed to validate and expand upon these findings. We found that a machine learning approach with SVM using the feature set adjective (JJ) + adverb (RB) + verb (VB) + noun (NN) + subjective words (SW) achieved the best performance (93.7%). Future studies would also benefit from the use of a more rigorous holdout evaluation dataset as opposed to the current relatively small dataset which was analyzed in this experiment using 10-fold cross validation.

Previous studies have also demonstrated that LDA is an effective method for finding keywords for topics. Although LDA did not perform the best out of the three methods, for this dataset it required only a small amount of words (150 words) to achieve relatively high accuracy. One reason for this is that we have a very small dataset (699 short statements) and limited topics mapping to ACGME competencies (6 topics). LDA is more robust for larger datasets and can assign multiple topics for one statement. With this small dataset, we often could not obtain more than one candidate keyword for a topic. For example, we only obtained 340 keywords for 500 topics in LDA. Features of adjective (JJ) + noun (NN) resulted in some improvements, but use of POS required more than three times the amount of words with LDA. IG is a robust feature filter in that it helps to improve computational performance by using only half the amount of words and all the POS types. However, while IG reduced features for adjective (JJ) + noun (NN), it did not help to improve topic performance. This indicates that selecting POS tags only as a feature is not effective for topic analysis. With our best IG model, all precisions were over 70% except SBP due to its small sample set (50 sentences). The machine learning approach SVM with IG feature selection achieved the best overall precision and recall of 76.3%. We also tried merging features from both LDA and IG, but this had a relatively lower level of overall performance. One final problem of the dataset we found is that about 100 sentences did not map to a competency or could map to multiple topics. We envision the use more sophisticated analysis with multiple topics in the analysis with a larger dataset and the application of semantic similarity to this problem, by empirically grouping or applying semantic similarity measures as a means to discover semantically similar terms for topic analysis[36].

We also found that it was hard for annotators to reach consensus with topic analysis for some of the statements. For example, one annotator labeled the sentence "She has supervised the interns in a caring manner." as ISC since the resident demonstrated the interpersonal and communication skills "as a member or leader of a health care team" (see Table 1). Another annotator thought that it belongs to SBP since the resident demonstrated "a responsiveness to the larger context and system of health care". Another statement, "very good clinical skills", was also marked differently by two annotators. One annotator thought it was a PC statement since it fit the definition of "residents must be able to provide patient care". Another annotator categorized it to MK because clinical skills are applications of medical knowledge to patient care. While this was resolved in this study with consensus between the two raters and discussion with a third rater, further work on this is needed.

To investigate the potential subtopics of each competency, we analyzed sentences in the largest competency in our dataset, ISC, and found several potential ISC subtopics. This indicates that a study on deeper topic classification may be valuable at the subtopic level within each competency. In this study we implemented topic analysis at an ACGME competency level, but future work could focus on different topic levels including greater detail in clinical skills, teaching, and leadership.

Further work in topic analysis could also help to add to the growing body of evidence examining the ability of evaluators to distinguish resident performance in each of the competency domains. Previous research in this area has used quantitative evaluation data. Studies with quantitative evaluation data have demonstrated that one or two underlying factors (clinical competence and interpersonal competence) were primarily being used to assess resident performance[37], suggesting that faculty evaluators are not distinguishing residents' performance in each of the competency domains. This, in turn, has sparked a debate in the literature as to whether the competencies can/should be assessed in isolation or whether we should be assessing the integration of the competencies[38]. Topic analysis allows us to potentially visualize comments around themes, which could then be analyzed more systematically in terms of the competency domain definitions. This might provide or refute other evidence that evaluators are not distinguishing performance in each of the competency domains.

In this system, topic analysis to classify evaluation texts along various ACGME competencies (see Table 1) and sentiment analysis to categorize evaluator attitudes (e.g. extremely positive) about the trainee were moderately successful in automatically categorizing trainee performance. As outcomes of the system, each evaluation statement was automatically assigned with both a sentiment and a competency. With this information more readily available, our vision is for medical educators to have better tools to more easily evaluate each individual resident along training competencies. Similar to other studies with quantitative evaluation data, our pilot data and automated sentiment analysis demonstrated that evaluators trend towards supportive comments, possibly at the expense of providing resident trainees with constructive criticism. The use of topic analysis had good face-validity and identified some potentially valuable keywords and themes, which as further developed, might be used in the development of future evaluations. Limitations and next steps include the need to expand this study to a larger set of evaluation text.

This work would also ultimately provide great practical value in understanding to what extent real world educational trainee outcomes such as resident probation, board scores, and quantitative evaluation ratings correlate with our automated analyses of qualitative assessments. Ultimately, it would be highly helpful and timely if these methods could be further developed to help instructors and program directors by providing early "signals" from these texts to identify trainees at risk for problems. This would allow for interventions to help trainees to be started near the beginning of training when the benefit and the potential for success is the greatest for trainees.

Our long-term vision is also to expand this work not only to correlate with resident outcomes, but also to analyze trainer (teaching) issues such as identifying redundancy in comments written by evaluators based off of techniques developed with our prior work[39, 40], which likely provide minimal meaningful feedback to trainees. In other words, we can estimate the attitude of trainers (teachers) to trainees (residents) based on variation degrees and sentiment degrees of evaluation texts. If a trainer always writes the same or very similar evaluation texts to all residents, we can detect this by using redundancy measurement methods previously developed. After optimizing our techniques for instructor evaluations of residents, we plan to expand our analysis to 360-degree ratings of the trainee, which are currently being performed by ancillary staff (e.g., nurses and social workers) and families. Further development of the system can also help generate evaluation summarization of each resident which would include quantitative evaluation data and objective outcomes such as standardized test scores and performance probation, as well as possibly to identify resident similarity based on evaluations and system performance. Ultimately, these techniques have the potential to help provide medical educators and medical education researchers with automated tools to synthesize qualitative evaluation data in a more streamlined and efficient manner.

**Acknowledgements**

The authors would like to thank the Departments of Pediatrics and Medicine and the Institute for Health Informatics at the University of Minnesota and American Surgical Association Foundation (GM) for support of this study.

**References**

1. Fletcher KE, Underwood W 3rd, Davis SQ, Mangrulkar RS, McMahon LF Jr, Saint S. Effects of work hour reduction on residents' lives: a systematic review. J Am Med Assoc 2005;294(9):1088-100.
2. Bell RH. Surgical council on resident education: a new organization devoted to graduate surgical education. J Am Coll Surg 2007;204(3):341-6.
3. Williams RG, Klamen DL, White CB, Petrusa E, Fincher RM, Whitfield CF, Shatzer JH, McCarty T, Miller BM. Tracking development of clinical reasoning ability across five medical schools using a progress test. Acad Med 2011;86(9):1148-54.
4. ACGME. http://www.acgme.org/acWebsite/home/home.asp.
5. Iglehart JK. Medicare, graduate medical education, and new policy directions. N Engl J Med 2008;359(6):643-50.
6. Fletcher KE, Reed DA, Arora VM. Patient safety, resident education and resident well-being following implementation of the 2003 ACGME duty hour rules. J Gen Intern Med 2011;26(8):907-19.
7. Goldstein MJ, Kim E, Widmann WD, Hardy MA. A 360 degrees evaluation of a night-float system for general surgery: a response to mandated work-hours reduction. Curr Surg 2004;61(5):445-51.
8. Edwards FD, Frey KA. The future of residency education: implementing a competency-based educational model. Fam Med 2007;39(2):116-25.
9. Baker K. Determining resident clinical performance: getting beyond the noise. Anesthesiology 2011;115(4):862-78.
10. Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: a review of the literature. Med Teach 2004;26(4):366-73.
11. Colbert CY, Ogden PE, Ownby AR, Bowe C. Systems-based practice in graduate medical education: systems thinking as the missing foundational construct. Teach Learn Med 2011;23(2):179-85.

12. Weinberger SE, Pereira AG, Iobst WF, Mechaber AJ, Bronze MS; Aliance for Academic. Competency-based education and training in internal medicine. Ann Intern Med 2010;153(11):751-6.
13. Baglia J, Foster E, Dostal J, Keister D, Biery N, Larson D. Generating developmentally appropriate competency assessment at a family medicine residency. Fam Med 2011;43(2):90-8.
14. Finn A, Kushmerick N. Learning to classify documents according to genre. J Am Soc Inform Sci Tech 2006;57(11):1506-18.
15. Kamps J, Marx M, Mokken RJ, De Rijke M. Using WordNet to measure semantic orientation of adjectives. Proc Int Conf Lang Res Eval 2004;5:1115-8.
16. Wiebe J. Learning subjective adjectives from corpora. Proc Nati Conf Arti Intel 2000:735-40.
17. Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proc Annu Meet Assoc Comp Ling 2000:417-24.
18. Benamara F, Cesarano C, Picariello A, Reforgiato D, Subrahmanian VS. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. Proc Int Conf Web Soc Med 2007:203-6.
19. Subrahmanian VS, Reforgiato D. AVA: Adjective-verb-adverb combinations for sentiment analysis. IEEE Intell Syst 2008;23(4):43-50.
20. Mulder M, Nijholt A, Den Uyl M, Terpstra P. A lexical grammatical implementation of affect. Lect Notes Comp Sci. 2004;3206:171-7.
21. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proc Conf Empi Meth Natu Lang 2002:79-86.
22. Boiy E, Hens P, Deschacht K, Moens MF. Automatic sentiment analysis of on-line text. Proc Int Conf Elec Pub 2007:349-60.
23. Dave K, Lawrence S, Pennock DM. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proc Int Conf WWW 2003:519-28.
24. Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping. Proc Conf Natu Lang Learn 2003:25-32.
25. Lin CY, Hovy E. The automated acquisition of topic signatures for text summarization. Proc Conf Comp Ling 2000:495-501.
26. Dumais ST, Furnas GW, Landauer TK, Deenvester S. Using latent semantic analysis to improve information retrieval. Proc Conf Hum Fac Comp Sys 1988:281-5.
27. Blei DM, Ng AY, Jordan MI, Lafferty J. Latent Dirichlet allocation. J Mach Learn Res 3 2003:993-1022.
28. New Innovations RMS Evaluations application. http://www.new-innov.com/pub/rms/main.aspx.
29. The Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml.
30. MPQA Subjective Lexicon. http://www.cs.pitt.edu/mpqa/subj_lexicon.html.
31. Weka. http://www.cs.waikato.ac.nz/ml/weka.
32. Stopword list. http://www.textfixer.com/resources/common- english-words.txt.
33. LVG. http://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_004.htm.
34. Stanford Topic Model Toolkit. http://www-nlp.stanford.edu/software/tmt.
35. Hunt RJ. Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. J Dent Res. 1986;65(2):128-30.
36. McInnes BT, Pedersen T, Pakhomov SV. UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. AMIA Ann Symp Proc 2009:431-5.
37. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. Acad Med 2009;84(3):301-9.
38. Epstein RM, Hundert EM. Defining and assessing professional competence. J Am Med Assoc 2002;287(2):226-35.
39. Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. AMIA Ann Symp Proc 2011:1612-20.
40. Zhang R, Pakhomov S, Melton GB. Automated identification of relevant new information in clinical narrative. Proc ACM SIGHIT Int Health Inform Symp 2012:837-41.