Modeling Doctor-Patient Communication with Affective Text Analysis

Taylan Sen, Mohammad Rafayet Ali, Mohammed (Ehsan) Hoque
Department of Computer Science
University of Rochester, United States
Email: [tsen, mali7, mehoque]@cs.rochester.edu

Ronald Epstein, Paul Duberstein
School of Medicine and Dentistry
University of Rochester, United States
Email: [ronald_epstein, paul_duberstein]@urmc.rochester.edu

***Abstract*— We present a method of automatic analysis of doctor-patient communication and present findings after applying this methodology in a post hoc study of communication between oncologists and their cancer patients (N=122). We analyzed several features of each participant in the conversation including the number of words spoken, the average positive/negative sentiment expressed, the number of questions asked, and the word diversity (unique word count). We found that the number of words spoken by the doctor is correlated with the highest doctor communication ability ratings made by patients. We additionally found that unsupervised clustering of conversation features into "styles" identified that certain styles are associated with higher communication ratings. Two well-defined styles emerged when clustering based on doctor word diversity and doctor sentiment: a high word diversity-neutral sentiment style, which was associated with higher ratings, and a low word diversity-positive sentiment style with lower average ratings. Machine learning models were trained to automatically predict whether a doctor-patient interaction will be rated high or not with a best-performing 71% test set accuracy.**

1. Introduction

Have you ever not understood your doctor but were either too shy, too confused, or just didn't know what to ask to get better clarification? Most people have at least yearly contact with their physician, the CDC estimating in 2016 that 83% of adults had contact with a health care professional [1]. In the US, the number of physician office visits in 2016 was 922.6 million [2]. Considering that effective doctor-patient communication has been correlated with better health outcomes [2], [3], [4], [5], a substantial opportunity exists to improve patient outcomes by improving doctor-patient communication. However, one major problem is: *What are the characteristics of an interaction which makes good doctor-patient communication?*

One of the most vulnerable patient populations is late stage cancer patients. Cancer caused an estimated 595,000 deaths in 2016 [8]. According to the National Cancer Institute, approximately 39.6% of the population will be diagnosed with cancer at some point in his/her lifetime [8], and nearly half will die of their disease [9]. Effective communication with cancer patients is especially important due to the emotional impact, the need to share complex information about the disease, prognosis and options, and the need for patient active involvement in complex decision-making. Furthermore, distressed patients may have compromised the ability to assimilate complex information and to participate as fully as they might wish in their care. In this paper, we developed an automatic communication analysis methodology which was applied in a post hoc analysis of the transcript data of 122 visits between cancer patients and their oncologists, and the associated post-visit patient survey ratings of the physician's communication behaviors [10]. The data for each visit includes a post-visit patient survey, an audio recording of the visit, and a text transcription of the audio recording.

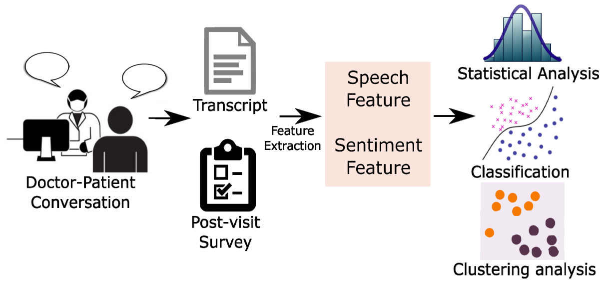


Figure 1. Overview of Communication Analysis

Based on the work of [11], [12], [13], [14], we extracted 14 communication features for each of the participants including sentiment, a number of words spoken, word diversity, and several other count features described in detail in the following sections. An exploratory data analysis was conducted in light of patient survey ratings regarding oncologist communication ability. A number of features were identified which correlated with higher communication survey scores. Specifically, the features of doctor word diversity and doctor spoken word count showed statistically significant differences between the group of doctors receiving the highest possible ratings and the other doctors.

We applied two machine learning models, including logistic regression and k-Nearest Neighbors, in order to predict from the conversation features whether an oncologist is in the class of those receiving the highest communication ratings or not. We found the k-Nearest Neighbor classification model to perform best, with an average accuracy of 71% in cross-validated test sets (AUROC=0.73). Good performance appears to be limited by the high variability in the feature values across each of the conversations.

Conversations were clustered using unsupervised learning to see if there were distinct styles among subsets of the extracted features. Several feature combinations were found to cluster into a low number of groups as the ideal number of groupings. When comparing the average ratings for each of these clusters, some clusters were found to have statistically significant differences in communication ratings. For example, doctors who were grouped into a cluster who used unique words and neutral sentiment had higher communication ratings compared to the cluster characterized by

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

fewer unique words and a more positive doctor sentiment.

One surprising finding was that while doctor sentiment alone showed no statistically significant difference between the highest performing doctors and other doctors when combined with doctor word diversity, it enabled better identification of good communication than was possible than with any feature alone.

Our contributions are summarized as follows:

•Identification of speech features which are correlated with communication outcomes including number of doctor words spoken and doctor word diversity;

•Development of an automatic analysis methodology to predict communication outcome from the speech transcripts; and

•Identification of common styles of doctor-patient communication through unsupervised clustering which is associated with better and worse communication outcomes.

2. Related Work

Many studies have found that effective doctor-patient communication is correlated with better health outcomes. Epstein and Street suggested that better communication can reduce suffering in cancer patients [7]. A 2001 Institute of Medicine report identified that patient-centeredness is one of the key ways to improve healthcare in the US [15]. Additional evidence supporting this position was gathered by Kaplan et al., their study finding that doctor-patient communication is consistently related to both physiological and behavioral metrics of health [3]. Stewart, et al., further found that communication is related to positive health [4]. The degree of patient-centeredness in physician-patient communication was found to affect patient satisfaction [16] and verbal behaviors were found to be associated with health outcomes [5]. Laws, et al., found some correlation between patient-centered communication metrics and conversation time [12].

A number of studies applied automated methodologies in the analysis of doctor-patient communication. Lacson, et al. developed a system for automatically analyzing and summarizing spoken medical dialogue, finding that doctors were able to answer questions regarding patient care from the automatically generated summaries [17]. Angus et al. developed a graphical visualization tool to model physician-patient dialogue to identify patterns of engagement between individuals including communication accommodation, engagement, and repetition [18]. Models of doctor-patient communication were also developed in several studies. Wallace, et al., developed a system based on a conditional random field model for automatically annotating topics in physician-patient communications [19]. More recently, Wallace, Dahabreh, and Trikalinos developed a speech model in which physician-patient conversations were clustered based on speech act features of the conversations and the resulting groupings correlated with patient survey ratings of physician communication [20]. Their technique further enabled a comparison of communication styles between the resulting groups of doctor-patient dyads, finding that advising patients without their permission is negatively correlated with patient satisfaction.

While the previous work advances understanding of effective doctor-patient communication, we are aware of no attempts to examine the combination of both affective and non-affective speech features automatically gathered in a machine learning framework. We propose a novel methodology to identify latent styles in doctor-patient communication regarding a number of affective and non-affective speech features. This approach was applied on a dataset of 122 doctor-patient conversations of late-stage cancer patients to link combinations of speech features with better patient-reported communication outcomes.

3. Study Data

3.1. Raw Data

Data gathered includes transcripts of physician-patient visits, audio recordings of the interactions, and patient survey results. All of the patients had stage 3 or stage 4 (late stage) advanced solid tumors. The recorded and transcribed interaction was for a routinely scheduled visit between the oncologist, patient, and in most cases, a family caregiver. Others may have been present (e.g. nurses as well). Data was gathered with a total of 40 oncologists and 122 patients. Patients completed Likert-type scales regarding their well-being and the physician's communication skill. The latter included:

1. My cancer doctor encouraged me to ask questions
2. My cancer doctor was willing to discuss any topic of importance to me
3. My cancer doctor gave me information I could understand
4. I felt understood by my cancer doctor
5. During this visit, do you feel that you got enough relevant information from your doctor?

Transcription of the recorded audio files was done by professional transcribers. In summary, the data for each patient that we analyzed includes a text transcription of the visit's audio recording and a post-visit survey of the patient's ratings of aspects of the doctor's communication skill. Three outliers were removed in which it appeared the patients misunderstood the direction of the Likert rating scales.

3.1. Extracted Features

Listed in Table I are the 14 features which were extracted from the doctor-patient conversation transcript files. Included are the speech features separately calculated for the doctor and patient of total number and percentage of words spoken in the conversation, the number of unique words spoken ("word diversity"), and the number of questions asked. The affective features used include the doctor and patient's positive, negative, neutral, and composite sentiment. Sentiment is generally understood to represent the associated positive or negative emotion. We extracted sentiment features using the well-known VADER (Valence Aware Dictionary for sEntiment Reasoning) tool [23] for sentiment analysis with the NLTK natural language processing toolkit [24]. In summary, VADER calculates sentiment through the use of a rule-based model which uses a sentiment lexicon (dictionary of words containing an associated valence measure). The sentiment lexicon used by VADER was produced from a human-labeled corpus in which humans rated sentiment as the overall positive, neutral, or negative emotion associated with a given word in a phrase or sentence. While the

TABLE I. EXTRACTED TRANSCRIPT FEATURES

Speech Features	Affective Features
# and % of words spoken by D	D composite sentiment
# and % words were spoken by P	D positive sentiment
# questions asked by D	D neutral sentiment
# questions asked by P	P composite sentiment
# of unique words spoken by D	P positive sentiment
# of unique words spoken by P	P neutral sentiment

sentiment features do not have a concrete definition, they are a collective result of a large number of human raters' understanding of positive and negative emotion associated with words. Examples of sentiment feature values for specific text will follow in the analysis of section 4.3. We hypothesized that word count would be a good indicator of the volume of knowledge/information communicated back and forth between the doctor and patient. This notion was supported by the findings of Laws et. al, in identifying a link between doctor-patient visit time and measures of patient-centeredness [12]. We further conjectured that the unique word count could identify either good or bad characteristic of communication. Ratzan identified that good communication in the health care setting is not embodied simply by repetition [21]. Alternatively, it is reasonable that through repetition, comprehension and successful information transfer would be more successful. Prior research has analyzed the number of doctor questions asked as well as the encouragement of patients to ask questions in relation to measures of effective communication [14]. We expected that the number of questions asked by the patient would correlate with patient-reported oncologist communication quality. Additionally, physician guidelines have suggested the importance of positive language in doctor-patient communication [22]. Several standard text analysis paradigms were not used that were not automated. For example, speech act coding methodology was not used because it generally requires human coding for accurate results.

4. Analysis

We first conducted a statistical analysis of the extracted features. A number of machine learning models were then applied to automatically predict patient-reported oncologist communication quality. We then used unsupervised clustering of the features to identify where distinct communication styles existed and analyzed whether different styles were associated with different communication effectiveness outcomes.

4.1. Statistical Analysis of Features

Shown in Figure 2a-e are the distributions of the survey ratings for each of the survey questions. As shown, the distributions are highly skewed, a vast majority of ratings being the highest score. Out of all the questions, 78% of the ratings were a “1”. Using a regression (continuous output) model to predict the scores would be inappropriate due to the highly skewed nature of the data. We thus summed the scores for all the questions as shown in Figure 2f, and

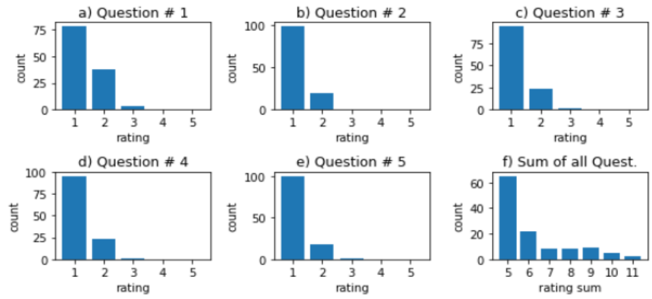


Figure 2. Histograms of the number of conversations receiving a given rating: a-e) based on a question number, f) sum of all questions

created two classes of conversations: those with maximum scores on all items assessing patient-reported oncologist communication quality and those who have less than the maximum score. By summing the scores on each question into a cumulative score, we obtain two advantages. First, the cumulative score reduces measurement error. Second, it is possible to partition the samples into two groups which are approximately equally sized. We labeled the resulting groups: A) those conversations in which the doctors

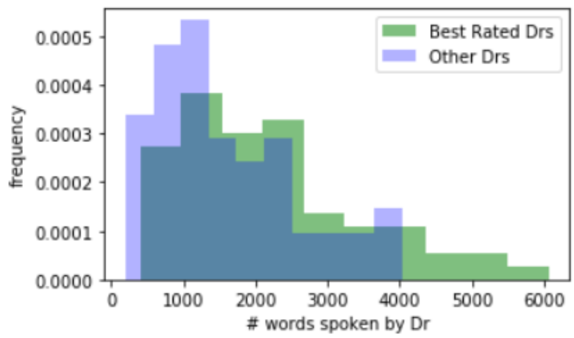


Figure 3. Comparative Histograms of the # Words Spoken by Doctor for Best Rated Doctors and Other Doctors

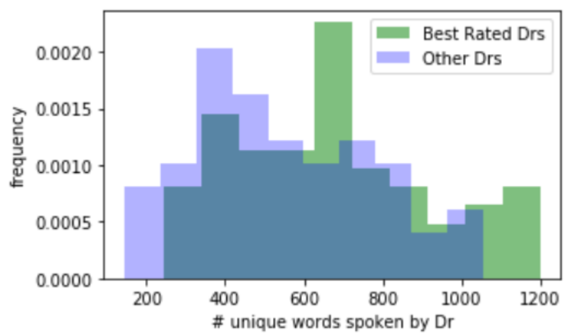


Figure 4. Comparative Histograms of the Doctor Unique Word Count for Best Rated Doctors and Other Doctors

TABLE II. STATISTICAL COMPARISON OF FEATURE AVERAGES BETWEEN BEST RATED AND OTHER INTERACTIONS

Feature	t-test p-value	Effect size Cohen's *d*	Best Rated Drs.	Other Drs.
# words spoken by doctor	**0.0034***	**0.542**	**2193**	**1578**
% words spoken by doctor	0.143	0.283	55.37	50.67
# words spoken by patient	0.053	0.358	1205	947
% words spoken by patient	0.870	0.031	31.98	31.54
Doctor composite sentiment	0.576	-0.104	0.48	0.48
Doctor neutral sentiment	0.352	0.174	0.55	0.54
Doctor positive sentiment	0.331	-0.182	0.41	0.42
Patient composite sentiment	0.825	-0.041	0.12	0.12
Patient neutral sentiment	0.887	-0.027	0.72	0.72
Patient positive sentiment	0.931	0.016	0.22	0.22
Doctor unique word count	**0.004**	**0.535**	**673**	**544**
Patient unique word count	0.063	0.347	446.0	383.5
# questions asked by doctor	0.531	0.116	39.12	36.78
# questions asked by patient	0.867	0.031	15.43	15.04

received the best possible rating ("Best Rated Doctors", N=65) and all other conversations ("Other Doctors", N=57).

The features collected were compared between the two resulting groups. Table II shows the means of each feature for each of the two groups, along with the uncorrected p-values of a Students t-test comparison, and the Cohen's *d* effect size. The most prominent differences (in terms of t-test p-value and Cohen's *d* effect size) were found in the features of *# doctor words spoken* and the *doctor unique word count*. Surprisingly, the number of questions asked by the patients did not show a substantial difference between the two groups. Additionally, none of the sentiment measures individually showed a substantial difference between the two classes. Shown in Figures 3 and 4 are the histograms of the number of words and number of unique words spoken by doctors and the in the Best Rated Doctors and the Other Doctors groups. As can be recognized readily from Figure 3, a significant portion of the Best Rated Doctors had word counts above 4000, compared to the Other Doctors who had no doctors with counts above 4000. Additionally, as shown in Table II, the Best Rated Doctors had a larger mean than the Other Doctors, the means being 2193 and 1578 respectively with an effect size of 0.542 (Cohen's *d*). Because we investigated 14 different features, it is important to consider the Bonferroni corrected p-value, which is 0.0476 (14 x 0.0034) for the *# doctor words spoken* feature.

Because the distributions diverge somewhat from a normal distribution, which is an assumption of the t-test, a Mann-Whitney-Wilcoxon (MWW) test was also performed (which does not assume any distribution type). The MWW test compares the medians instead of the means of each group. The medians of the two groups are 2024 and 1328 with a p-value of 0.00327 uncorrected and 0.0458 with Bonferroni correction for 14 features, suggesting that the two groups indeed have different medians. The findings that the best communication ratings are correlated with higher *# doctor word counts* is consistent with prior research [12].

We obtained similar findings for the *doctor unique word count* feature as shown in Figure 4. However, the Bonferroni-corrected t-test p-value for doctor unique word count is not < 0.05. (t-test p=0.061, MWW p=0.056). It is possible that the larger # of unique words spoken by the Best Rated Doctors is a result of either explaining more different information, or explaining the same amount of information in more ways rather than solely through repetition. It is important to note that the results only show correlation and do not show causation. However, it is conceivable that by speaking more, doctors are more likely to transfer important information more completely and fully address patient questions, resulting in better patient-reported oncologist communication quality.

4.2. Predicting Survey Responses with Logistic Regression and k-Nearest Neighbor Classifiers

Both logistic regression and k-Nearest Neighbor (KNN) models were applied to the extracted features for each interaction to try to predict the interaction's classification among Best Rated Doctors vs. Other Doctors. Both of these models were implemented with the scikit-learn machine learning toolkit [25]. Five-fold cross-validation was used to evaluate model performance. We ran a logistic regression using all features with L1 regularization validation set optimized hyper-parameters. We additionally applied logistic regression without regularization using subsets of up to 5 of the features. The KNN model was also trained on both all the features, as well on subsets of up to 5 of the features. For each KNN model, the ideal k was selected by picking the k between 2 and 70 with the best performance on a validation set.

The best performing model was KNN with k=13 yielding a mean test set accuracy of 0.71. The most common features used among the cross-validated KNN models were *#/% of doctor words*, *#/% patient words*, and *# unique doctor words*. The average AUROC of the 5-fold cross-validated KNN models is 0.73, indicating mediocre prediction performance. In summary, classification is not able to accurately distinguish conversations with the best communication ratings, yielding an accuracy of only 71%. However, the classifier does provide the ability to automatically analyze a conversation, and flag those which have a higher probability of exhibiting poor communication.

4.3. Clustering

In order to see whether the features naturally formed clusters, i.e., styles of communication, we performed an unsupervised clustering analysis. Unlike classification, unsupervised clustering allows us to identify styles independent of ratings information. This is important because, whereas machine learning partitions the feature space in a way that maximizes the classification accuracy, clustering instead forms clusters by minimizing intra-cluster feature difference. In order to identify if such styles existed, we applied k-means clustering [26][25] on normalized features to find clusters of conversation features. The resulting clusters were analyzed to see if they had differences in their ratings. Here, we present only those results with the k which gave the highest Silhouette coefficient [27][25], and for which the post survey ratings showed substantial differences between the clusters. We used the Silhouette coefficient in order to avoid manually fixing the number of clusters, instead of using an objective measure of the ideal k (number of clusters). Clustering was performed on all pairs of the features. For several pairs of the features, we found significant differences in the communication ratings between the clusters. In all of these cases, the ideal number of clusters was determined to be 2. The pairs of features in these cases are listed in TABLE III. The middle column of TABLE III shows the resulting clusters visually. The right column of TABLE III shows the average ratings received by the two clusters for each of the five post-visit survey questions. Additionally, a Student's t-test comparison was performed on the two clusters for each of the five post-visit survey questions. Some of the resulting p-values were less than 0.05. Due to the exploratory nature of this investigation, Bonferroni correction was not applied to these values in order to prevent type I errors and thus they do not represent statistical significance in this section. Differences with $p<0.05$ between the two clusters are marked by a star (*) in TABLE III.

1) % words spoken by doctor and Doctor positive sentiment

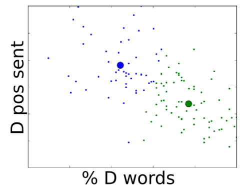
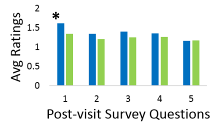
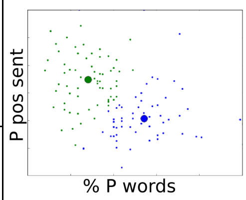
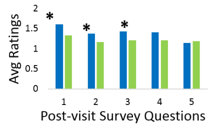
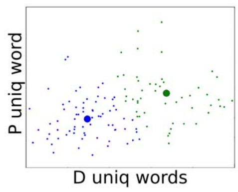
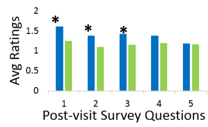
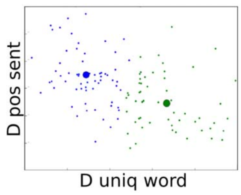
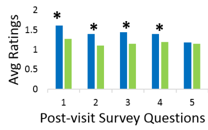
Clustering resulted in two clusters, a blue cluster and green cluster as shown in TABLE III-A, each with differing feature levels. The green cluster is characterized by doctors with a higher word percentage (> 60%) and less positive sentiment, the blue cluster with doctors with lower word percentage and more positive sentiment.

We noticed that for the post-visit survey question 1 (Q1,"My cancer doctor encouraged me to ask questions.") the ratings were notably different between the two clusters. The doctors who spoke more and kept a lower positive sentiment (green cluster), were, according to patient ratings, more likely to encourage the patient to ask questions. On the other hand, the doctors who participated in the conversation equally (i.e., word %) and had more positive sentiment (blue cluster) were rated to be less likely to encourage the patient to ask questions. This finding might represent situations in which doctors offer reassurance (a positive sentiment) without understanding the patient's experience (lower word count) are less effective communicators, as noted in prior work [28], suggesting that trust and empathy lead patients to rate physicians' communication as more effective. For hypothetical example: Dr. A: "I think everything will be fine." [# words=6, positive sentiment = 0.27] vs Dr. B: "I see that you are still concerned, and I'll make sure that we leave no stone unturned." [# words = 17, positive sentiment = 0.12]

2) % words spoken by patient and Patient positive sentiment

As shown in TABLE III-B, we found notable differences in three post-visit survey question ratings when we clustered based on *% words spoken by patient* and *Patient positive sentiment*. One cluster (green cluster) ended up having a higher positive patient sentiment (~0.7) and a lower percentage of words spoken by the patient (<40%). The other cluster (blue cluster) has less positive patient sentiment (~0.4) and an interaction in which the patient and doctor spoke an equal % (i.e., % words spoken by patient ~50%). On average, patients felt more encouraged to ask questions (Q1), felt their doctor was willing to discuss any topic (Q2), and felt the doctor's provided information was understandable (Q3) when the patients maintained a positive sentiment and spoke less (green cluster). These findings are consistent with the clustering results in Table-I. Specifically, when patients use fewer words (lower word percentage), that usually indicates the doctors are using more words (higher word percentage). In summary, patients who have a positive disposition and talk less are more likely to feel they had effective communication with their doctor. Further research would be necessary to determine whether this pattern represents a) good care of an informed and involved patient who is content with his/her role in decision-making or b) avoidance of difficult topics and maintaining unrealistic optimism in the context of serious illness.

TABLE III. RESULTS OF CLUSTERING.

<table>
<tr><th>Feature Pair</th><th>Clusters</th><th>Average ratings of each cluster on each question</th></tr>
<tr><td>A.
% words spoken by doctor
Doctor positive sentiment</td><td>
</td><td>
</td></tr>
<tr><td>B.
% words spoken by patient
Patient positive sentiment</td><td>
</td><td>
</td></tr>
<tr><td>C.
Doctor Unique Word Count
Patient Unique Word Count</td><td>
</td><td>
</td></tr>
<tr><td>D.
Doctor Unique Word Count
Doctor positive sentiment</td><td>
</td><td>
</td></tr>
</table>

3) Doctor Unique Word Count and Patient Unique Word Count

TABLE III-C shows that when clustering based on the *doctor unique word count* and the *patient unique word count* the resulting two clusters have a significant difference in the ratings of the same questions (Q1, Q2, Q3) as shown in Table 3-2. The doctors and patients who used more unique words end up in one cluster (green). The other cluster (blue) contains those interactions where both doctors and patients used fewer unique words. When doctors and patients used more unique-words, the patients felt that they are being encouraged to ask more questions, they are being informed about their future, and they received understandable information. The t-test p-values associated with each of these questions are found to be < 0.05. This might represent a subset of doctors whose explanations were clearer as a result of using unique words instead of repeating themselves.

4) # of unique words spoken by doctor and Doctor positive sentiment

Shown in Table III-D are the clustering results with the features of *# of unique words spoken by doctor* and *Doctor positive sentiment*. The doctors who used more unique-words with less positive sentiment formed one style (green cluster). The other style automatically identified is characterized by doctors who used fewer unique words and have more positive sentiment (blue cluster). We found these two groups' ratings were significantly different for four of the post-visit survey questions (Q1, Q2,Q3, Q4). For these four questions, the green cluster received higher ratings than the blue cluster. From these results, it appears that the patients prefer the clear explanation of their situations from their doctors. It is also evident that having positive sentiments does not guarantee a higher rating from the patient, similar to findings noted in section 4.3.1)

4.4. Linguistic Inquiry and Word Count

After observing the differences in doctor and patient word counts between the Best Rated Doctors and the Other Doctors as well as observing the importance of these features in our predictive classifier and clustering analysis, we conducted a Linguistic Inquiry Word Count (LIWC) analysis of the transcripts [29]. Listed in TABLE IV are the LIWC word categories with the lowest 10 and bottom 3 t-test p-values and effect sizes for their difference between the Best Rated Doctors group and the Other Doctors group. Note that these p-values are not Bonferroni corrected. The categories are listed in order of their strongest effect size. The largest effect size was for the *You*, *I*, and *Personal* word categories, suggesting that the relational components of clinical care, in particularly patient-centered behaviors on the part of the physician, is interpreted by patients as effective communication. In contrast, more population-based or experimental-based language would be found to be less patient-centered and effective. Consider the hypothetical example of a difficult decision in the face of progressing cancer along with hypothetical responses: Dr. C, using a less personal, more population-oriented, and less affiliative approach: *"Studies have shown that this chemotherapy regimen leads to a 5-10% reduction in mortality at 6 months but with the potential for severe side effects, which are more common with comorbidities such as yours. Based on those data, the best approach is not using chemotherapy at this time. Importantly, there are things that can be done to improve symptoms, quality of life and comfort, which we will put in place."* [# you/I = 1, # words = 70, positive sentiment = 0.19,] vs Dr. B, using a more patient-centered, personal, and pronoun-rich approach: *"I have looked at your situation. There are statistics about how well this chemotherapy regimen works, which in general is modest. But I see your situation as different from other patients in several ways, which makes me think that your risk of having a bad reaction to the chemotherapy might be even higher than most. I know this is disappointing for you and me, but I think that the risks would clearly outweigh the benefits. Of course, I am always committed to working hard to help you live a fully and as comfortably with whatever time you have left. "* [you/I = 12, # words = 99, positive sentiment = 0.17]

TABLE IV. DIFFERENCES IN LIWC BETWEEN BEST RATED DOCTORS AND OTHER DOCTORS

LIWC Category	p-value	effect size	effect size Cohen's d
You	0.001	0.64	-0.64
I	0.002	0.58	-0.58
Personal	0.002	0.57	-0.57
Total	0.003	0.56	-0.56
Perceptual	0.003	0.54	-0.54
Social	0.004	0.54	-0.54
Exclusion	0.004	0.53	-0.53
Time	0.004	0.53	-0.53
Impersonal	0.004	0.53	-0.53
Adverbs	0.004	0.53	-0.53
….	…	…	…
death@Death	0.807	0.04	-0.04
Friends	0.869	-0.03	0.03
Anxiety	0.980	0.00	0.00

5. Conclusion

In this paper, we investigated the usefulness of affective text analysis in modeling doctor-patient communication skills and found that although several features had differences among the best-rated Doctors and other Doctors We found that traditional machine learning classification methods struggle to automatically identify the Doctors with the highest communication ratings with a reasonable accuracy. However, it was found that unsupervised clustering was capable of identifying several communication styles associated with higher and lower communication ratings.

We hope to apply the presented model as a way of automatically flagging doctor-patient communications which may benefit from human review. While this was an exploratory study, we hope to investigate the causality of the primary findings that the number of words spoken by the Doctors is correlated with good communication and that adopting certain speaking styles has a positive effect on communication ratings and potentially patient outcomes.

References

[1] Centers for Disease Control and Prevention, "Summary Health Statistics : National Health Interview Survey , 2015 Table A-18a , page 1 of 9 Summary Health Statistics : National Health Interview Survey , 2015 Table A-18a , page 2 of 9," 2015.

[2] Centers for Disease Control and Prevention, "National Ambulatory Medical Care Survey: 2012 State and National Summary Tables," pp. 1–40, 2015.

[3] S. H. Kaplan, S. Greenfield, and J. E. Ware, "Kaplan, Assessing_the_Effects_of_Physician_Patient, Medical Care, 1989.pdf," *Med. Care*, vol. 27, no. 3 Suppl, pp. S110-27, 1989.

[4] M. Stewart, J. Brown, A. Donner, I. McWhinney, and J. Oates, "The Impact of Patient-Centered Care on Outcomes : The Journal of Family Practice," *J. Fam. Pract.*, vol. 49, no. 9, pp. 796–804, 2000.

[5] R. S. Beck, R. Daughtridge, and P. D. Sloane, "Physician-patient communication in the primary care office: a systematic review," *J. Am. Board Fam. Pract.*, vol. 15, no. 1, pp. 25–38, 2002.

[6] Institute of Medicine & Committee on Quality of Health Care in America, *Crossing the Quality Chasm: A New Health System for the 21st Century*, vol. 323, no. 7322. 2001.

[7] R. M. Epstein and R. L. Street Jr., "Patient-Centered Communication in Cancer Care: Promoting Healing and Reducing Suffering," *Communication*, p. 222, 2007.

[8] N. C. Institute, "Cancer Statistics." [Online]. Available: https://www.cancer.gov/about-cancer/understanding/statistics.

[9] N. C. Institute, "Cancer Stat Facts: Cancer of Any Site." [Online]. Available: https://seer.cancer.gov/statfacts/html/all.html.

[10] M. Hoerger *et al.*, "Values and options in cancer care (VOICE): study design and rationale for a patient-centered communication and decision-making intervention for physicians, patients with advanced cancer, and their caregivers.," *BMC Cancer*, vol. 13, p. 188, 2013.

[11] B. C. Wallace, I. J. Dahabreh, T. A. Trikalinos, M. B. Laws, I. Wilson, and E. Charniak, "Identifying Differences in Physician Communication Styles with a Log-Linear Transition Component Model," no. Laws 2013, pp. 1314–1320, 2014.

[12] M. B. Laws, L. Epstein, Y. Lee, W. Rogers, M. C. Beach, and I. B. Wilson, "The association of visit length and measures of patient-centered communication in HIV care: A mixed methods study," *Patient Educ. Couns.*, vol. 85, no. 3, pp. 1–15, 2011.

[13] J. H. Kahn *et al.*, "Linked references are available on JSTOR for this article : Measuring emotional expression with the Linguistic Inquiry and Word Count," vol. 120, no. 2, pp. 263–286, 2016.

[14] I. Hazen-Klemens and E. Lapinska, "Doctor-patient interaction, patient's health behavior and effects of treatment," *Soc. Sci. Med.*, vol. 19, no. I, pp. 9–18, 1984.

[15] a. Wolfe, "Institute of Medicine Report: Crossing the Quality Chasm: A New Health Care System for the 21st Century," *Policy, Polit. Nurs. Pract.*, vol. 2, no. 3, pp. 233–235, 2001.

[16] P. Little *et al.*, "Observational study of effect of patient centredness and positive approach on outcomes of general practice consultations," *Bmj*, vol. 323, no. 7318, pp. 908–911, 2001.

[17] R. C. Lacson, R. Barzilay, and W. J. Long, "Automatic analysis of medical dialogue in the home hemodialysis domain: Structure induction and summarization," *J. Biomed. Inform.*, vol. 39, no. 5, pp. 541–555, 2006.

[18] D. Angus, B. Watson, A. Smith, C. Gallois, and J. Wiles, "Visualising conversation structure across time: Insights into effective doctor-patient consultations," *PLoS One*, vol. 7, no. 6, 2012.

[19] B. C. Wallace, M. B. Laws, K. Small, I. B. Wilson, and T. A. Trikalinos, "Automatically annotating topics in transcripts of patient-provider interactions via machine learning.," *Med. Decis. Making*, vol. 34, no. 4, pp. 503–12, 2014.

[20] T. Althoff, K. Clark, and J. Leskovec, "Identifying differences in physician communication styles with a log-linear transition component model," *Proc. Natl. Conf. Artif. Intell.*, vol. 2, no. January, pp. 1314–1320, 2014.

[21] S. C. Ratzan, "Health literacy: communication for the public good.," *Health Promot. Int.*, vol. 16, no. 2, pp. 207–214, 2001.

[22] C. Zinn, "Doctors told to use positive language in managing pain," *BMJ*, vol. 326, p. 301, 2003.

[23] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Eighth Int. AAAI Conf. Weblogs …*, pp. 216–225, 2014.

[24] S. Bird, S. Bird, and E. Loper, "NLTK : The natural language toolkit NLTK : The Natural Language Toolkit," *Proc. COLING/ACL Interact. Present. Sess.*, no. March, pp. 69–72, 2016.

[25] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[26] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[27] P. J. Rousseeuw, "Silhouettes:a graphical aid to the

interpretation and validation of cluster analysis," vol. 20, pp. 53–65, 1987.

[28] R. M. Epstein, T. Hadee, J. Carroll, S. C. Meldrum, J. Lardner, and C. G. Shields, "'Could this be something serious?': Reassurance, uncertainty, and empathy in response to patients' expressions of worry," *J. Gen. Intern. Med.*, vol. 22, no. 12, pp. 1731–1739, 2007.

[29] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin," *Development*, vol. 1, no. 2, pp. 1–22, 2007.