Understanding patient satisfaction with received healthcare services: A natural language processing approach

Kristina Doing-Harris, PhD[1], Danielle L. Mowery, PhD[3], Chrissy Daniels, MS[2], Wendy W. Chapman, PhD[3], Mike Conway, PhD[3]
[1]Westminster College, Salt Lake City, UT;
[2]Director of Strategic Initiatives, University of Utah, Salt Lake City, UT;
[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

Important information is encoded in free-text patient comments. We determine the most common topics in patient comments, design automatic topic classifiers, identify comments' sentiment, and find new topics in negative comments. Our annotation scheme consisted of 28 topics, with positive and negative sentiment. Within those 28 topics, the seven most frequent accounted for 63% of annotations. For automated topic classification, we developed vocabulary-based and Naïve Bayes' classifiers. For sentiment analysis, another Naïve Bayes' classifier was used. Finally, we used topic modeling to search for unexpected topics within negative comments. The seven most common topics were appointment access, appointment wait, empathy, explanation, friendliness, practice environment, and overall experience. The best F-measures from our classifier were 0.52(NB), 0.57(NB), 0.36(Vocab), 0.74(NB), 0.40(NB), and 0.44(Vocab), respectively. F-scores ranged from 0.16 to 0.74. The sentiment classification F-score was 0.84. Negative comment topic modeling revealed complaints about appointment access, appointment wait, and time spent with physician.

Introduction

Patient satisfaction

Patient satisfaction ratings can be a good indicator of clinical effectiveness and patient safety[1]. However, the free-text comment fields, which are filled out in nearly 50% of patient surveys, are underutilized[2]. The Center for Medicare and Medicaid Services (CMS) and Agency for Healthcare Research and Quality (AHRQ) developed a national standard for reporting patient satisfaction called the *Hospital Consumer Assessment of Healthcare Providers and Systems* (HCAHPS). Siegrist et al.[2] report that hospital participation in HCAHPS is very high. In fact, the number of surveys collected is higher than other customer satisfaction surveys. The level of patient response indicates that there is an enormous amount of free-text information available. In 2014, the University of Utah collected 105,000 free-text patient satisfaction comments. In order to understand the causes for patient dissatisfaction, quality improvement abstractors review every patient-generated comment, which is both labor-intensive and expensive. Large-scale automated or semi-automated review would save time and money. It would also facilitate scaling the analysis of these comments for benchmarking over time.

Analyzing Free-text comments

The information available in free-text comments has been identified using qualitative methods[1,3,4]. Lopez, et al.[3] developed a complex taxonomy of patient comments (Table 1). It includes global themes of *overall excellence, negative sentiment*, and *professionalism*. They also identified specific factors, for example *interpersonal manner*, *technical competence*, and *system issues*. Doyle, et al.[1] echoed Lopez' et al.'s topic categories when they identify search terms for a meta-analysis of patient experience. They divide the terms into relational aspects and functional aspects. Relational aspects are equivalent to *interpersonal manner*. They include *emotional and psychological support*, *patient-centered decisions*, *clear information*, and *transparency*. Functional aspects are equivalent to Lopez et al.'s *professionalism*, *technical competence*, and *systems issues*. Functional aspects include *effective treatment*, *expertise*, *clean environment*, and *coordination of care*.

In terms of sentiment analysis, using a qualitative methodology, Lopez et al.[3] categorized 712 online reviews of physicians from the websites ratemds.com and Yelp.com according to polarity and learned that most reviews were rated positive (63%) recommending a patient's physician to others. Another study by Ellimoottil et al. reviewed physician rating sites for the scores for 500 urologists[5]. The free-text comments were classified as extremely positive, positive, neutral, negative and extremely negative. They found that most ratings (75%) were extremely positive, positive, or neutral. A meta-analysis of six physician rating websites also found that around 70% of ratings were positive[6].

Table 1: **Lopez[3] taxonomy of patient satisfaction themes**

Overall Excellence – Recommendation	*Negative Sentiment – Intent not to Return*	*Professionalism*
	SPECIFIC FACTORS	
Interpersonal Manner	**Technical Competence**	**System Issues**
Empathic	Knowledgeable	Appointment Access
Friendly	Detailed	Appointment Wait Time
Helpful	Efficient	Practice Environment
Trustworthy	Clinical Skills	Practice Health IT
Time Spent During Appointment	Follow-up	Practice Location
Put at Ease	Referrals	Cost of Care
Listens	Perceived Poor Decision Making	Negative View of Healthcare
Explains	Perceived Successfulness of Treatment	Method of Physician Selection
Longevity of Relationship with Clinician	Complementary-Alternative Medicine	

Natural Language Processing

Manually analyzing hundreds of thousands of free-text comments is difficult. Technology such as Natural language processing (NLP) can be used to make it more manageable. NLP has been used to classify comments by topics (**topic classification**)[7], to encode the polarity of sentiment expressed within a comment i.e., positive or negative (**sentiment analysis**)[8-11], and to determine if comments include unforeseen topics (**topic modeling**)[4,12]. Greaves et al.[7] used topic classification to classify 6,412 free-text online comments about hospitals from the English National Health Service. They employed three topics – *overall recommendation*, *cleanliness*, and *treatment with dignity.* They created a Naïve Bayes multinomial classifier. It achieved F-measures of 0.89 (*overall recommendation*), 0.84 (*cleanliness*), and 0.85 (*treatment with dignity*).

There has been a lot of work in sentiment analysis both inside and outside of the patient satisfaction domain. In the SemEval-2016 task 4, sentiment analysis on Twitter, the winning team for the two-class (positive, negative) task used a combination of convolutional neural networks, topic modeling, and word embeddings generated via word2vec. They achieved an F-score of 0.80 and an accuracy of 0.86[8]. Other investigators used keywords to indicate sentiment.[9,10] These papers did not verify the accuracy of their vocabulary-based assessment. They relied on an accuracy measure of 75% for the dictionaries they employed. In a head-to-head comparison between commercial and non-commercial sentiment analysis tools for classifying healthcare survey data, Georgiou, et al.[11] found that the WEKA implementation of Naïve Bayes' performed the best, with a weighted F-measure of 0.81.

For topic modeling, the Brody et al.[4] study, mentioned above, applied Latent Dirichlet Allocation (LDA) to 33,654 online reviews of 12,898 New York-based medical practitioners. Their model identified words associated with both specialty-independent themes (e.g., *recommendation, manner, anecdotal, attention, scheduling*) and specialty-specific themes (e.g., general practitioner: *prescription and tests*, dentist: *costs*, obstetrician/gynecologist: *pregnancy*). Maramba, et al.[12] analyzed a free-text response from a post-consultation postal survey using a modified version of the English GP Patient Survey (GPPS) questionnaire. The question asked for any further comment. 3,462 individual comments were collected. They separated patient comments based on their overall rating of their experience on a 5-point Likert scale. *Very satisfied* and *fairly satisfied* were grouped as satisfied. *Very dissatisfied* and *fairly dissatisfied* were grouped as *dissatisfied*. The words "surgery", "excellent", "service", "good", and "helpful" were the five most distinctive words from satisfied patients, while the words "doctor", "feel ", "appointment", "rude", and "symptoms" were the five most distinctive words in the comments from dissatisfied patients.

Our current work builds on these previous efforts by addressing patient satisfaction from Press Ganey patient satisfaction survey data gathered from a health care system rather than publicly accessible online reviews and NHS surveys. Additionally, we developed an NLP-powered classifier that combines vocabulary-based and machine learning-based methods to analyze both the topic and sentiment of each free-text comment. For the long-term goal of this project, we aim to leverage NLP methods to automatically analyze free-text fields in Press Ganey patient surveys at the University of Utah hospital system in order to streamline quality improvement efforts e.g., trending historical patient experience data, helping direct future quality improvement efforts, and acquiring a better understanding of patient experiences as these relate to patient outcomes more generally. Our short-term goals are to determine common topics of patient satisfaction and dissatisfaction from free-text, patient survey comments, to create an NLP solution to automatically annotate comments with these topics, to analyze these comments for their polarity, and to identify sub-topics described within these comments to assist quality improvement efforts.

Methods

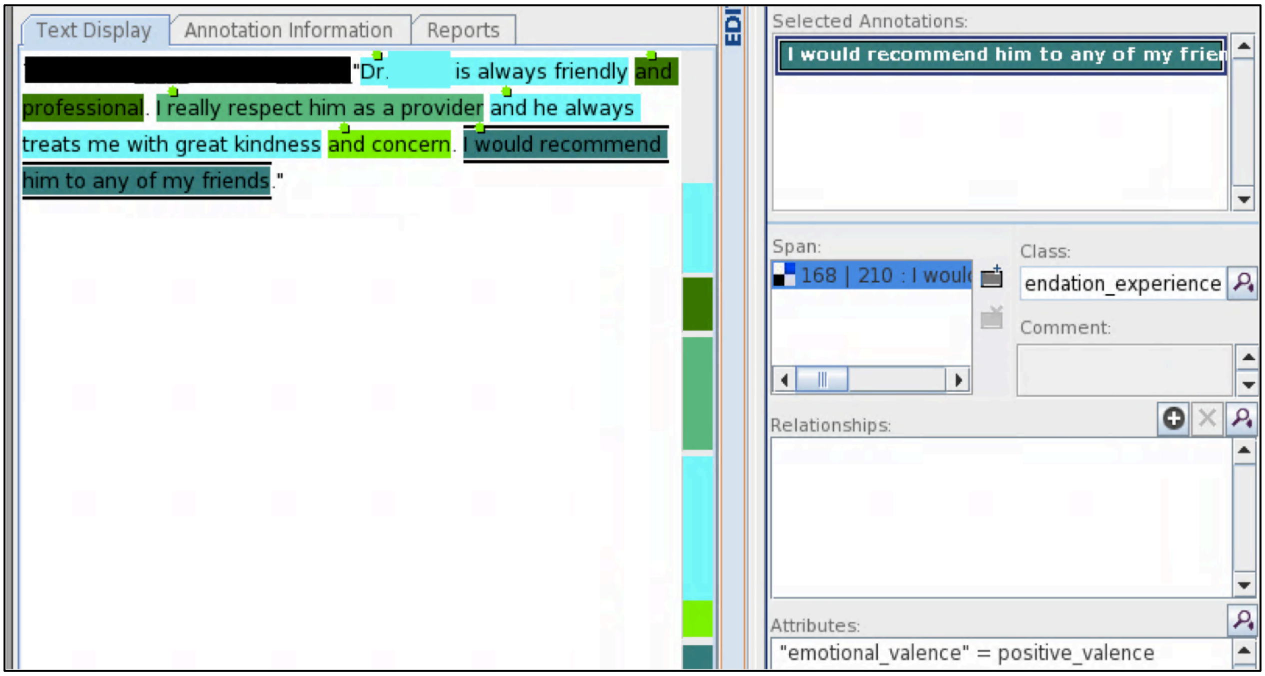
In this IRB-approved study (IRB_00081172), we obtained the 51,234 Press Ganey patient satisfaction responses from the University of Utah Health Care System (UUHCS) that were generated between January 1, 2014 and December 31, 2014. First, we developed a schema for characterizing topics from patient survey responses (Table 2) and validated our schema with an annotation study. Next, we trained two supervised classifiers (one a vocabulary-based classifier and one a Naive Bayes' classifier) to automatically tag responses with topics from the schema. Then, we identified patient's emotional valence toward these topics, using a separate trained classifier. Finally, we used LDA to cluster terms associated with negative experiences in an attempt to learn new topics.

Table 2. Annotation categories developed for this project.

Advice_experience	Helpful_experience	Practice_environment_experience
Appointment_access_experience	Intent_not_to_return	Practice_family_friendliness_experience
Appointment_wait_experience	Intent_to_return	Professional_experience
Clinical_skill_experience	Knowledge_of_patient_experience	Recommendation_experience
Decision_making_experience	Knowledgeability_experience	Relationship_longevity_experience
Efficiency_experience	Listened_to_experience	Time_spent_experience
Empathy_experience	MyChart_experience	Treatment_success_experience
Explanation_experience	Overall_experience	Trustworthy_experience
Follow_up_experience	Patient_autonomy_experience	
Friendly_experience	Percieved_bias_experience	

Annotation Study

We developed an annotation schema for topics and sentiment recorded in free-text satisfaction responses. Starting with the taxonomy by Lopez et al.[3] in Table 1, we added and removed categories in consultation with the Exceptional Patient Experience team (author CD) from UUHSC. In total, we created the 28 annotation categories listed in Table 2. Any group of words could be annotated separately with as many categories as were appropriate. Each annotation was also given a sentiment of positive, negative or neutral, an example is given in Figure 1.



[redacted] "Dr. is always friendly and professional. I really respect him as a provider and he always treats me with great kindness and concern. I would recommend him to any of my friends."

Selected Annotations: I would recommend him to any of my frie

Span: 168 | 210 : I woul

Class: endation_experience

Comment:

Relationships:

Attributes: "emotional_valence" = positive_valence

Figure 1. An example of an annotated file.

A set of 300 documents was randomly selected from the dataset of patient survey free-text responses for manual annotation. Annotators A1 and A2 developed the annotation guidelines on batches of 100 documents until agreement was reached using consensus review. We considered inter-annotator agreement (IAA), F1-score, to have reached an acceptable level when the overall agreement, as reported by eHOST, was 0.74. A third annotator (A3 who has experience in the hospital quality domain) was trained in the application of the annotation schema. The annotations generated by the third annotator were compared to the adjudicated set. With our initial annotation scheme, we quickly found that sufficient IAA for all topics could not be reliably maintained. Therefore, we focused on the 7 most common (*overall, appointment access, appointment wait, explanation, friendliness, practice environment, and empathy*) with positive and negative sentiment, which represented 63% of all annotations.

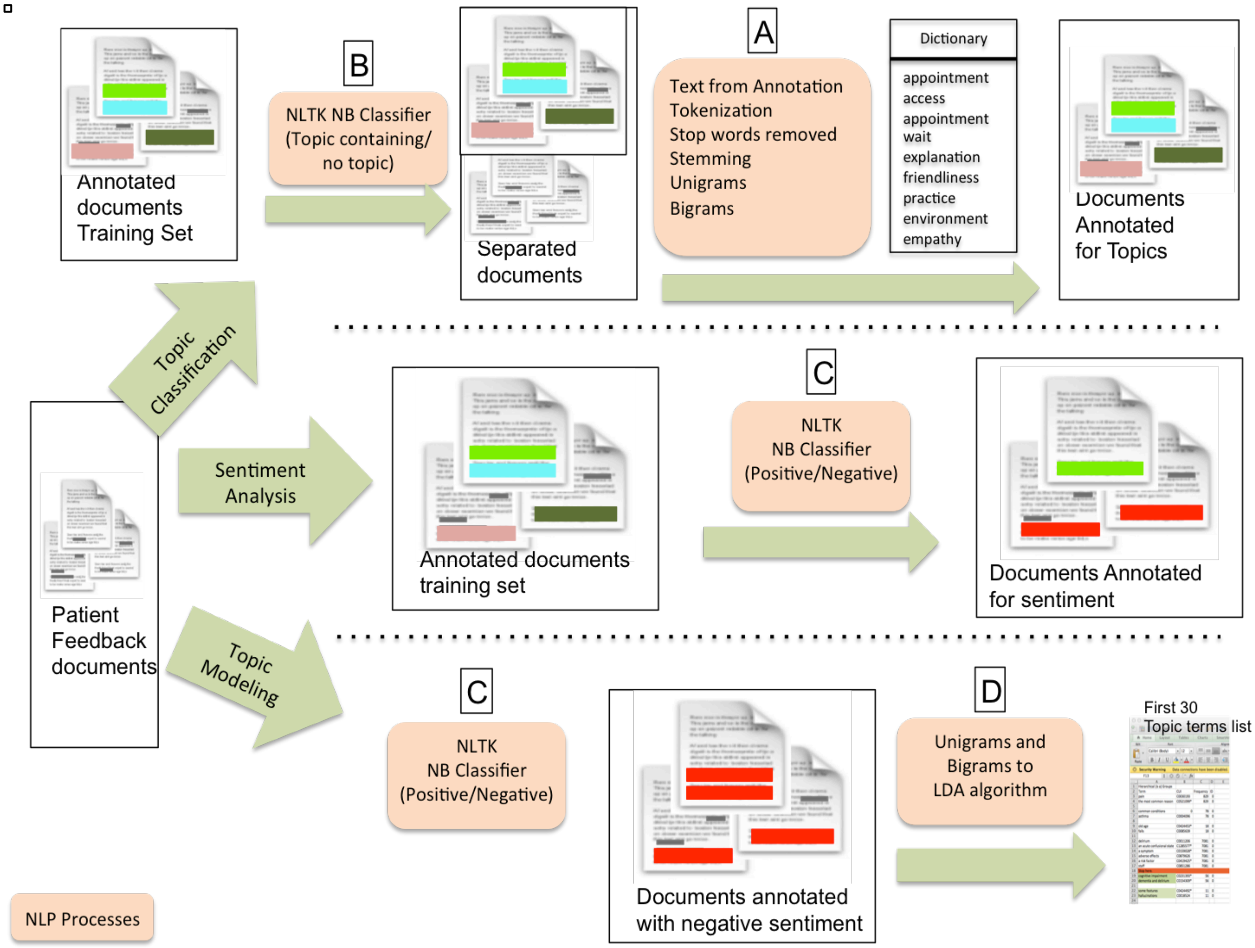

Annotated documents Training Set
B
NLTK NB Classifier (Topic containing/ no topic)
Separated documents
A
Text from Annotation
Tokenization
Stop words removed
Stemming
Unigrams
Bigrams
Dictionary
appointment access
appointment wait
explanation
friendliness
practice environment
empathy
Documents Annotated for Topics
Topic Classification
Patient Feedback documents
Sentiment Analysis
Annotated documents training set
C
NLTK NB Classifier (Positive/Negative)
Documents Annotated for sentiment
Topic Modeling
C
NLTK NB Classifier (Positive/Negative)
Documents annotated with negative sentiment
D
Unigrams and Bigrams to LDA algorithm
First 30 Topic terms list
NLP Processes

Figure 2. Chart illustrating the flow of Patient feedback documents through our processing systems.

Topic classification

For topic classification, we developed both vocabulary-based and machine learning-based approaches. Our approach is illustrated in the top line of Figure 2. For the vocabulary-based method (our baseline approach), we used the annotated topics in the 300 adjudicated documents to generate a vocabulary. We gathered the text from each topic and used the Natural Language Tool Kit (NLTK) for Python[*] to tokenize the comments, remove non-alphabetic characters, stem (Snowball Stemmer) the resulting tokens, and remove stop words from each tokenized comment. We separated the results into n-grams: unigrams, e.g., “terrific,” and bigrams, e.g., “terrific service” (process A in Figure 2). The lists of n-grams were compared so that each n-gram appeared on only one list. If an n-gram was on more than one list, the list with the highest frequency of its occurrence got to retain it. The top five n-grams for each topic are listed in Table 3. From Table 3, we observed that the top n-grams associated with *appointment access* and *appointment wait* are associated with time. In contrast, *practice environment* n-grams are often associated with cleanliness and temperature. *Empathy* and *friendliness* n-grams describe feelings and service actions.

[*] http://www.nltk.org/

We compared two methods for classifying comments into each topic. First, we used a simple dictionary look up approach. For each unigram and bigram in a comment, we found the corresponding topic. All matched topics were retained because many comments had more than one hand annotation. Like all attempts at document classification there was a tradeoff between identifying the document category correctly (i.e., precision) and finding all the documents that belonged in that category (i.e., recall). We observed that allowing n-grams to appear in the lists of more than one topic increased recall at the cost of precision.

Table 3. The five most prevalent unigrams and bigrams for each topic category.

Topic	**Type**	**N-gram feature set**
Overall	Unigrams	fantast, awsom, satisfi, absolut, fabul
	Bigrams	good experi, great experi, excel experi, excel servic, far good
Appointment access	Unigrams	cancel, week, holiday, apart, saturday
	Bigrams	schedul appoint, get appoint, abl get, get see, could get
Appointment wait	Unigrams	hr, end, period, realiz, paperwork
	Bigrams	wait time, time minut, wait hour, long wait, exam room
Explanation	Unigrams	detail, futur, describ, bring, comdit
	Bigrams	answer question, explain everyth, explain thing, explain would, happen futur
Friendliness	Unigrams	courteous, polit, interact, courtesi, paper
	Bigrams	staff friend, alway friend, nurs friend, realli nice, feel like
Empathy	Unigrams	compassion, sensit, respect, encourag, situat
	Bigrams	show concern, realli care, made feel, feel like, wait time
Practice environment	Unigrams	clean, wash, confirm, equip, thermomet
	Bigrams	wash hand, alway clean, wait area, thermomet probe, hot drink

We first thought to increase precision by creating a dictionary listing for n-grams, which indicate the comment does not contain any of the topics of interest. By creating a list for comments that were "not annotated," meaning that they did not contain any of the topics we were interested in. However, looking at our low recall scores, further striping n-grams would be counter productive because it would cause more comments to be missed. So we trained a Naïve Bayes (NB) classifier using the 300 adjudicated documents to classify comments as "topic containing" or "no topic" (process B in Figure 2) before we used the vocabulary-based system and learned n-grams (keywords) on the "topic containing" comments. Otherwise, the system could always find at least one topic matched word in every comment leading to very poor precision scores.

For comparison against the vocabulary-based approach, we trained a machine learning approach leveraging the Naïve Bayes (NB) algorithm. We tested decision tree and SVM models as well, but we found the best performance with Naïve Bayes'. We used the NLTK (i.e., featx) methods to reduce the text to lowercase, stem words to their lemma, group the stemmed words into n-grams, and convert the n-grams into feature vectors. We used the NLTK NB classifier, with default settings. We divided the data set into training/test data sets (75%/25%), one binarized set per annotated topic, each set was balanced between "topic containing" and "no topic" comments (i.e., positive and negative examples of the class). This process is not pictured in Figure 2, in order to retain readability. We applied the trained algorithm to the blind test set. We report the results on the 25% test data compared to the vocabulary approach for the seven most common topics. Performance is reported as precision, recall and F-measure.

Sentiment Analysis

Using the same training and test sets and encoded n-gram features from the topic classification, we developed a classifier to categorize a comment based on its sentiment. In line with the head-to-head comparison performed by Georgiou, et al.[11] and our results with the "topic containing"/"no topic" classification, we chose to again use Naïve Bayes'. Specifically, we trained the NLTK NB approach to classify comments based on sentiment categories of positive or negative (process C in Figure 2). We report the performance for each sentiment category on the test set.

We then ran both the trained and tested NB topic classifier and NB sentiment classifier on the remaining roughly 50,000 patient satisfaction comments. We report the distribution of *positive* and *negative* comments by topic class.

Topic Modeling

To complement the topics annotated through manual review, we completed an unsupervised topic modeling study. We first classified the full 51,234 comments with sentiment categories using the NB sentiment classifier (process C in Figure 2, repeated in the figure for clarity). For all comments classified as *negative*, we provided the unigrams and bigrams to an LDA algorithm with a preset maximum of 30 topics using the gensim package[*]. We report the n-grams associated with 10 of the 30 topics learned by the algorithm and if a topic suggests one of our seven most common topics, we also provide a topic label (process D in Figure 2).

Results

Annotation Study

In total, we annotated 1,374 documents consisting of 2,021 annotations. All three annotators annotated the same 300 documents in 100 document sets. Figure 3 illustrates how IAA changed across the three annotators and 3 document sets. In general, there was little improvement between sets 1 and 2. Some categories' performance even decreased indicating that new comments were difficult to categorize. Between sets 2 and 3 there is a general increase in IAA indicating that annotators had reached a common understanding. *Empathy* and *practice environment* categories, however, still demonstrate a lack of agreement.

Annotators A1 and A2 discussed categories and re-annotated sets 1 and 2 until they reached an understanding. On set 3 their agreement was 0.74 overall as reported by eHOST. At this point the most frequent topics (*overall*, *appointment access*, *appointment wait*, *explanation*, and *friendliness)* had high agreement levels of (0.83, 0.91, 0.79, 1.00, 0.76, respectively)[13]. *Empathy* (0.47) and *practice environment* (0.44) were much lower.

The same procedure of discussion and re-annotation of the first 2 sets was repeated with A3. His overall agreement with the adjudicated document set from A1 and A2 was reported by eHOST to be 0.73 overall. Agreement between A3 and the adjudicated documents for the most frequent categories was 0.77, 0.83, 0.77, 0.86, and 0.77, respectively. *Empathy* (0.67) and *practice environment* (0.60) demonstrated lower agreement, although not as low as the agreement between A1 and A2.

Taken across the 2 sets of IAA scores, the categories with the most agreement are *explanation* (1.00, 0.86), *appointment access* (0.91, 0.83), *overall* (0.83, 0.77), *appointment wait* (0.79, 0.77), *friendliness* (0.76, 0.77), *empathy* (0.47, 0.67) and *practice environment* (0.44, 0.60). Agreement with A3 is tied for the middle three. So the score between A1 and A2 is used to decide order.

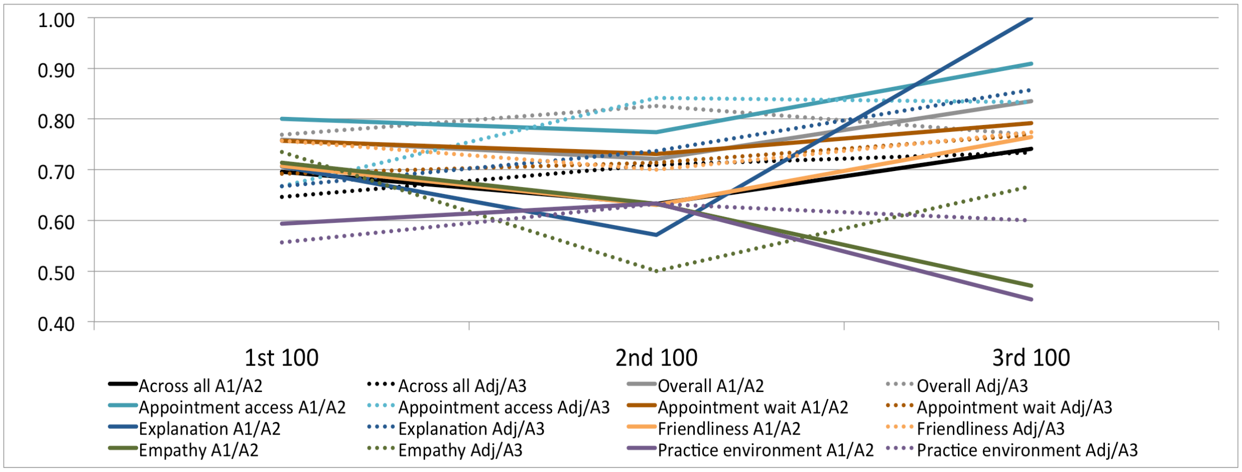


Figure 3. Inter-annotation scores across three document batches (each containing 100 documents). Agreement statistic used is F1.

Topic classification

While generating the vocabulary, we observed, for *overall experience* for example, 123 unigrams. Of those, 19 were repeats of previously occurring unigrams. We refer to the 104 n-grams, without the repeats, as unique. Since a

[*] https://radimrehurek.com/gensim/models/ldamodel.html

unigram could not appear in more than one list, the number of unique n-grams is the number of words that were available to identify topics. 58 bigrams (20 of them unique) were found for *overall experience*. Figure 4 shows the counts of unigrams and bigrams, with the number of unique in each case, across the seven most common topics. *Practice environment, appointment access,* and *appointment wait* generated the most n-grams and the most unique n-grams (Figure 4). Looking ahead at the classification results in Table 3, there is no obvious relationship between the number of unique n-grams and performance. *Explanation* had the fewest unique n-grams and the highest performance. *Overall experience* had the most unique n-grams was directly in the middle of the performance scores.

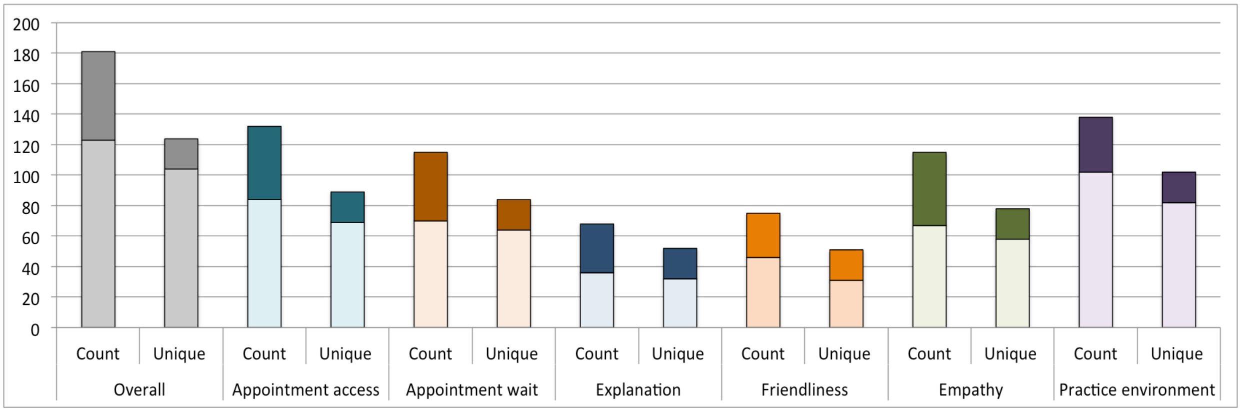


Figure 4. Total n-gram counts and unique n-gram counts for each topic. Unigrams = light color (bottom); Bigrams = dark color (top).

The vocabulary was extracted from the 300 adjudicated documents and tested on the remaining 1,074. Table 4 illustrates the results of the topic classification for the top seven topics. NB classification for *explanation* was good (0.74). Performance for the remaining 6 topics were fair, ranging from 0.36 – 0.57. Generally, precision was higher than recall for the vocabulary matching approach. In choosing the vocabulary, we opted for higher precision. However, this strategy was not successful for the more difficult categories of *practice environment* and *empathy*. These were the only topics for which the vocabulary matching outperformed the machine learning approach of applying NB. The precision for the NB *overall* classifier was high, but recall was lower. The topic classification results echo the IAA results, demonstrating the difficulty of the task. *Explanation* was the best performer in both places. *Empathy* and *practice environment* are at the bottom. *Friendliness* and *overall* do not follow this trend.

Table 4. Topic classification for n=1,074 annotations. Classifiers are listed by their highest F-measure. IAA indicates the rank order of the topic IAA scores. **Bold** indicates highest performance between approaches.

Topic	**IAA**	**Precision**		**Recall**		**F-measure**	
		Vocabulary	**Naïve Bayes**	**Vocabulary**	**Naïve Bayes**	**Vocabulary**	**Naïve Bayes**
Explanation	1	0.42	**0.90**	0.35	**0.63**	0.38	**0.74**
Appointment wait	4	0.31	**0.54**	0.32	**0.60**	0.31	**0.57**
Appointment access	2	**0.45**	0.44	0.46	**0.64**	0.46	**0.52**
Practice environment	7	**0.38**	0.24	**0.61**	0.57	**0.48**	0.34
Overall	3	0.66	**0.85**	**0.30**	0.27	**0.44**	0.41
Friendliness	5	**0.50**	0.32	0.31	**0.57**	0.39	**0.40**
Empathy	6	**0.41**	0.18	**0.33**	0.14	**0.36**	0.16

Sentiment Analysis

Separate vocabulary and NB classifiers were developed to analyze the sentiment of a comment. These classifiers used all 1,374 documents split randomly into 75% for training and 25% for testing. The algorithm performed with a precision of 0.90 and recall of 0.80 for *positive* sentiment, and a precision of 0.79 and recall of 0.90 for negative sentiment. The overall F-score of the system was 0.84.

Combining Topic and Sentiment

We applied the two NB classifiers trained on 1,374 documents. At this point we trained the model on 100% of the documents because we used the trained model to annotate the 49,860 un-annotated patient satisfaction comments.

The classifiers tagged the comments with 73,801 annotations. Overall patients have *positive* experiences. Specifically, *empathy, friendliness,* and *explanation* are more often *positive* experiences; in contrast, *appointment wait, appointment access,* and *practice environment* are more often negative experiences. The breakdown of polarities for the top 7 topic categories is illustrated in Figure 5.

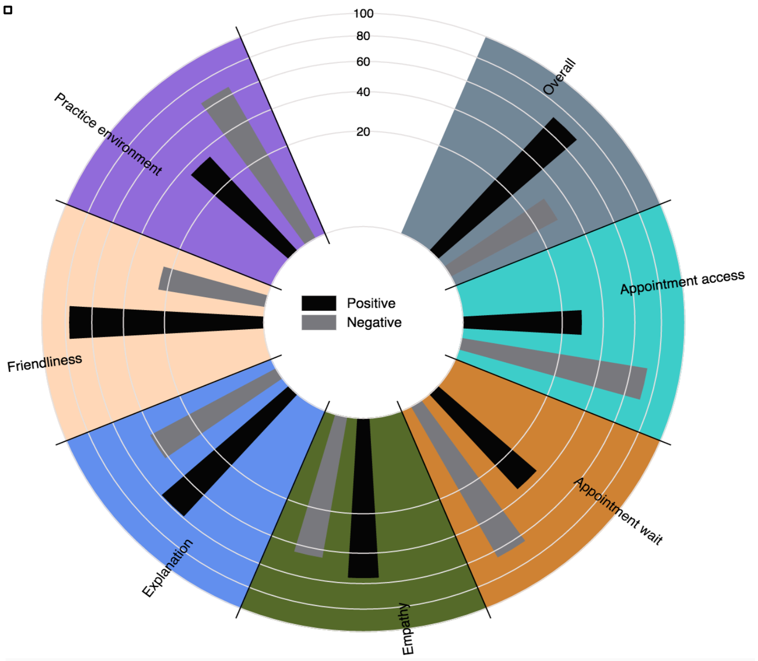


Figure 5. Distribution of positive and negative comments according to topic.

Topic Modeling

We applied a topic modeling algorithm to patient satisfaction comments from 2014 that had been automatically flagged as *negative* by our sentiment analysis tool to provide a snapshot of negative topics (Table 5). Topic models can be difficult to interpret[14], but it can be seen that most of the themes identified refer to discontent regarding the topics we named *appointment wait* and *appointment access*.

Table 5. Topic words associated with each label from 10 of the 30 topics learned.

Topic #	Topic words	Label
1	back, called, minutes, waited, come, doctor, behind, running, told, would, see, receptionist	appointment wait
2	time, patient, help, spent, bit, short, kind, enough, doctor, needed, amount, year	time spent with provider
3	one, another, first, saw, seeing, weeks, physician, eye, doctor, quickly, contact, would	appointment wait
4	wait, waiting, room, time, exam, delays, delay, minutes, informed, taken, times, area	appointment wait
5	felt, almost, like, pain, something, found, bad, brought, rushed, although, believe, scheduler	appointment access
6	clinic, call, office, phone, appointment, need, days, would, back, someone, got, center	appointment access
7	much, hours, two, wish, appreciate, seems, clinic, would, hard, question, convenient, sick	appointment wait
8	get, able, schedule, appointment, appt, always, lot, done, around, see, talk, trying	appointment access
9	appointment, visit, seen, scheduled, time, even, early, day, arrived, right, actually, though	appointment access
10	made, appointment, appointments, since, last, seemed, experience, available, month, years, months, helped	appointment access

Discussion

In our topic classification study, we found that the common topics of patient satisfaction and dissatisfaction from free-text, patient survey comments were *appointment access*, *appointment wait*, *empathy*, *explanation*, *friendliness*, *overall experience* and *practice environment*. In just over half of the cases (four out of seven) the NB classifier was more successful than the vocabulary-based system in classifying documents as belonging to the topic in question. Investigating comments for their sentiment, using an NB classifier over the entire document set, we found 70% to be *positive* and 30% to be *negative*. This breakdown in sentiment was also found in our manual annotations where 71% of the top 7 annotations and were *positive* and 75% of all annotations were *positive*. Using topic modeling to determine if there were topics we had not considered within the negative comments, we found that the topics described within the negative comments, reflect the common topics of *appointment access* and *appointment wait*.

Our initial plan was to create a topic classification system that would reflect the full complexity of the topics found using qualitative methods. We created an extensive taxonomy of topics, which could be annotated with reasonable overall agreement. However, IAA was only acceptable for five of the seven most common categories. *Empathy* and *practice environment*, although they were commonly used were not often agreed upon. In the case of *empathy*, it may be due to the difficulty in distinguishing between *empathy* and related topics, like *friendliness*. The relatively low agreement for the *practice environment* category is perhaps due to conflating overall *negative* experiences with issues relating to the *physical environment*.

The ability to create a high performing topic classifier may have been affected by the only fair agreement on annotations. The classifier with the best overall F-measure was *explanation*, which had the highest IAA. The worst performing classifier *empathy* was for the topic with one of the lowest agreement scores. The middle four NB classifiers did not show a relationship between F-measure and IAA rank. This pattern did not hold for the vocabulary-based systems. The best vocabulary system was for practice *environment*, which had the second lowest agreement rank. The vocabulary system with the lowest F-measure was for *appointment wait*, which had a middling agreement rank. We also found that as we added documents that had been annotated by only A3, the classifier performance decreased. It seems likely that A3 experienced drift in their annotations.

Low performance on topic classification, even with our restricted topic set, indicates that perhaps both approaches are unable to capture the semantic information in the patient comments. Performance may be improved by creating an ontology of the dictionary terms, a structure that reflects synonymy, hierarchical relationships between terms, and term function. For instance, “excellent” may be synonymous with “great”, in this context. “Really excellent” may be an intensified version of “excellent” and both “really excellent” and “excellent” can be used in combination with “environment” or “interactions.” The former would be associated with the topic *practice environment*, while the latter with *friendliness*. This kind of information may be encoded in convolutional neural networks, which have been successfully used in sentiment analysis[8].

By focusing on solely on those topics on which annotators can reach acceptable agreement, and conflating easily confused topics, we may be able to dramatically increase our classification performance. Perhaps combining the categories of *empathy*, *friendliness* and *helpfulness* (from our bigger list) we could create a topic that is easier to distinguish from the other topics. It is also possible that the number of annotations of each topic led to poor performance. However, the best performing classifier was for *explanation*, which had far fewer occurrences (92) than *overall* (587), which had one of the worst performing classifiers. Future work includes training on a larger set of data and using more semantic features e.g., encoding words from categories such as Time, Affect, Positive Emotion, and Negative Emotion, etc. from the Linguistic Inquiry Word Count lexicon[*].

The sentiment classifier performed very well, which is consistent with the larger training set due to a smaller number of classes. Overall system performance was above the threshold of 0.80. It performed within the range of Greaves, et al.[7,15,16] three-topic classifications e.g., 0.84 vs. 0.89, 0.84, and 0.85. Applying this classifier to our large dataset, we found 70% had positive sentiment compared to Lopez, et al.[3] who found 63% of online reviews had positive sentiment.

Good sentiment analysis performance allowed us to select the comments with *negative* sentiment for topic modeling. Our topic modeling results echo our topic classification results in that the predictive topic words indicate two of our most common topics *appointment access* and *appointment wait*. These two are also the topics with the most comments with *negative* sentiment. They are the only topics with more than 50 comments with *negative* sentiment across the annotated dataset. Brody et al.[4] also found *scheduling* to be a common review theme. We may have been more likely to find unexpected topics if we had looked at the least commonly occurring words in the negative comments. However, finding uncommon, unexpected topics may not be useful for quality control.

[*] http://liwc.wpengine.com/

As we have noted in this discussion, the main limitation of this work is the relatively poor performance of the topic classification and topic modeling systems. However, the problems encountered are common to NLP systems. In the case of topic classification, the more topics you have the more difficult the task becomes for both humans and machines. For topic modeling, the system's performance was not a problem. The results simply show that there are no unexpected topics found in the most frequent negative comments.

Conclusion

This study is a promising start to creating a method of analyzing free-text comments in patient satisfaction surveys that will help streamline patient satisfaction efforts e.g., trending historical patient experience data, helping direct future quality efforts, and acquiring a better understanding of patient experiences more generally. We are actively improving topic classification algorithm performance, expanding the granularity of sentiment classes to address strength of emotional valence (*strong positive, weak positive, neutral, weak negative*, and *strong negative* sentiment) and implementing monthly reports with statistics and visualizations to streamline patient satisfaction improvement efforts at UUHSC.

Acknowledgements

We would like to thank the anonymous reviewers for valuable comments. This work was partly funded by the University of Utah Healthcare System Hospital Project funds.

References

1. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013;3:e001570–0.
2. Siegrist RB Jr. Patient Satisfaction: History, Myths, and Misperceptions. Virtual Mentor American Medical Association Journal of Ethics 2013;15:982–7.
3. López A, Detz A, Ratanawongsa N, Sarkar U. What Patients Say About Their Doctors Online: A Qualitative Content Analysis. J Gen Intern Med 2012;27:685–92.
4. Brody S, Elhadad N. Detecting salient aspects in online reviews of health providers. AMIA Annu Symp Proc 2010;2010:202–6.
5. Ellimoottil C, Hart A, Greco K, Quek ML, Farooq A. Online Reviews of 500 Urologists. JURO 2013;189:2269–73.
6. Emmert M, Sander U, Pisch F. Eight questions about physician-rating websites: a systematic review. J Med Internet Res 2013;15:e24.
7. Greaves F, Millett C, Nuki P. England's Experience Incorporating 'Anecdotal' Reports From Consumers into Their National Reporting System: Lessons for the United States of What to Do or Not to Do? Medical Care Research and Review 2014;71:65S–80S.
8. Nakov P, Ritter A, Rosenthal S, Sebastiani F. SemEval-2016 task 4: Sentiment analysis in Twitter. Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US (forthcoming). 2016
9. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. PLoS ONE 2015;10:e0136341–10.
10. Song B, Lee C, Yoon B, Park Y. Diagnosing service quality using customer reviews: an index approach based on sentiment and gap analyses. Service Business 2015;:1–24.
11. Georgiou D, MacFarlane A, Russell-Rose T. Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. Science and Information Conference 2015;:352–61.
12. Maramba ID, Davey A, Elliott MN, Roberts M, Roland M, Brown F, et al. Web-Based Textual Analysis of Free-Text Patient Experience Comments From a Survey in Primary Care. JMIR Med Inform 2015;3:e20.
13. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960;20:37–46.
14. Chang J, Gerrish S, Wang C. Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems 2009;
15. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, et al. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf 2014;23:838–46.
16. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013;15:e239.