Research Report

Using Natural Language Processing to Code Patient Experience Narratives

Capabilities and Challenges

Daniel Ish, Andrew M. Parker, Osonde A. Osoba, Marc N. Elliott, Mark Schlesinger, Ron D. Hays, Rachel Grob, Dale Shaller, Steven C. Martino

For more information on this publication, visit www.rand.org/t/RRA628-1

Published by the RAND Corporation, Santa Monica, Calif.

www.rand.org

Preface

This research was conducted as part of the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) program. CAHPS is a research initiative of the Agency for Healthcare Research and Quality (AHRQ) to further knowledge on patient experience of care. One component of the CAHPS program advances tools that assess patients' experiences with health care and facilitate use of these products to improve quality of care. Since 1995, AHRQ has been awarding five-year cooperative agreements to private organizations to plan and conduct research for the CAHPS program. The RAND Corporation and the Yale School of Public Health are current grantees, with Westat providing program support under contract to AHRQ. Collectively, AHRQ and these three research organizations are known as the CAHPS Consortium.

This report is intended to be of interest to a broad audience, including health care organizations seeking to better understand and improve patient experience with care, policymakers providing guidance to providers, and academic researchers interested in the analysis of patient experience narrative data. As such, the report is written for audiences with a range of technical expertise and perspectives. Technical language is generally avoided, except where its absence could lead to lack of clarity or precision. When technical language is used, terms are initially defined using intuitive language.

This report specifically builds off of a technical advisory group meeting convened on October 29, 2019 by the CAHPS research team to develop principles and considerations for automated narrative analysis of patients' comments about their experiences with health care ("patient narratives"). The CAHPS research team conducted this demonstration study with these considerations in mind.

RAND Health Care, a division of RAND, promotes healthier societies by improving health care systems in the United States and other countries. RAND Health Care does this by providing health care decisionmakers, practitioners, and consumers with actionable, rigorous, objective evidence to support their most complex decisions.

The authors wish to thank Drs. Gondy Leroy and Bill Marcellino, whose suggestions helped improve the clarity of this report, and Dr. Jennifer Cerully, who organized and facilitated the technical advisory group that informed this research.

For more information, see www.rand.org/health-care, or contact

RAND Health Care Communications
1776 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
(310) 393-0411, ext. 7775
RAND_Health-Care@rand.org

Contents

Figures

Tables

Summary

Narratives about patients' experiences with health care contain a wealth of information about what is important to patients and are valuable in contextualizing and improving patient experience. Patients provide narrative descriptions of their experiences in a range of settings, including online communities, patient portals, and consumer rating sites. Here we focus on narratives collected from patients using standardized open-ended questions that appear on systematically sampled patient experience surveys. As illustrated in Box S.1, such narratives convey ample actionable information that (a) elaborates on existing domains covered by closed-ended questions contained in patient experience surveys and (b) incorporates additional domains valued by patients.

Box S.1. Illustrative Patient Narrative (Deidentified)

Dr. X is an outstanding physician. I feel very fortunate to be one of her patients. She is very thoughtful about my complex medical history and always takes very seriously any concerns that I express about my current health. She is extremely warm and caring and is also extremely smart. A very talented doctor! Once, I was experiencing an unusual symptom and Dr. X not only provided a thorough exam and ordered appropriate tests but she wouldn't let me leave the office until she had the test results. She went to great lengths to get those results because she knew I was quite concerned about what I was experiencing. It made me feel truly taken care of—I felt a level of trust and safety knowing that she would always go above and beyond to take care of me. In my experience, it is very unusual to find that level of personal commitment. I was already a huge fan of Dr. X, but that was an endearing experience that solidified my loyalty to her. Dr. X is very easy to talk. She has a good sense of humor and is very curious. I always feel well understood when we talk, and she takes my concerns very seriously. I can't say enough good things about Dr. X. She is wonderful.

The value of patient narratives for quality improvement is matched by the challenge of organizing, analyzing, and generating actionable insights from the data, especially when faced with many narratives. Rigorous qualitative analysis by human coders is likely infeasible for very large numbers of narratives. One potential solution to mitigate the resource requirements of human coding of narratives is the use of computer algorithms to extract structured meaning from unstructured natural language, referred to as *natural language processing* (NLP).

Although computer science and computational linguistics have been making advances in NLP since the 1950s, use of NLP in the realm of patient experiences with health care, as measured by standardized patient surveys, is a relatively new undertaking, with much of potential value still to be realized. *Our overarching objective in conducting this study was to provide an open-access demonstration of the feasibility of extracting key information from a large set of narratives on patients' experiences with ambulatory care using NLP and to highlight*

key challenges faced when doing so. We focused specifically on the scoring and organizing potential of NLP because we believe that it is crucial to provide health care consumers and providers access to verbatim narratives, tagged with and organized by content descriptors, along with useful summary information that the NLP analysis affords.

Our efforts revolved around the two steps that must be performed by a machine learning (ML) system designed to classify narratives into codes such as those typically applied by human coders (e.g., positive or negative statements regarding care coordination). These steps include numerically representing the text data (in this case, entire narratives as they are provided by patients, such as the one above) and then classifying the data by codes based on that representation. In doing so, we highlight a set of potential pitfalls, including *overfitting* the model (whereby as one adds more complexity to a model, one stops capturing generalizable lessons and starts capitalizing on the idiosyncrasies of the specific dataset), *imbalanced data* (and specifically trying to identify rarely mentioned aspects of care), and *insufficiently sized data sets* (which limit statistical power, especially with rare classes).

Data for this study came from the Massachusetts Health Quality Partners (MHQP) 2017 Patient Experience Survey, which asked adult patients to report their experiences with a specific primary care provider and with that provider's practice. The 2017 survey was based on the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Patient Centered Medical Home (PCMH) Survey developed by researchers from the National Committee for Quality Assurance (NCQA) and the Consumer Assessment of Healthcare Providers and Systems (CAHPS) Consortium. In all, we analyzed comments from 4,219 patients who received care from 780 practices. To construct a training sample from which the machine would learn, two human coders independently coded a subset of 1,490 of these comments. These coders assessed the presence or absence of 26 classifications such as positive or negative statements about patient-provider communication or care coordination.[1] Many of these classes (e.g., access to care) correspond to traditional numeric scores derived from closed-ended CAHPS survey questions, whereas others (e.g., emotional rapport) derive from prominent themes identified in past human coding of patient narratives. One set of classes pertains to a more complex theme—actionability—which represented a more challenging goal for automated coding, because in human-coding, it often cuts across elements of a narrative and involves multiple criteria.[2]

In a series of simple experiments, we used ML algorithms to predict human-generated codes applied to real-world patient narratives in order to explore the feasibility of using these methods in the patient narrative context. Because the size of the data set is modest, the performance of our

---

[1] The methodological field of human qualitative analysis tends to describe tagging text with *codes*, whereas the methodological field of ML tends to refer to *classes*, which in this case can be the presence or absence of a code. Here, we use the terms *code* and *class* interchangeably.

[2] A narrative is deemed "actionable" if it contains sufficiently detailed information about an experience that can plausibly be used to modify problematic health care practices and emphasize effective ones (Grob et al., 2019).

classifiers may not fully reflect those that could be produced by an effort involving a larger data set. We explored and compared four closely related approaches to deploying these algorithms. One approach is based on an older technique for developing numerical representations of text, the bag-of-words (BOW) approach, which is based solely on the frequency of the occurrence of words. The remaining three approaches utilized a more modern open-source natural language model called BERT (for "Bidirectional Encoder Representations from Transformers") to assign these numerical features. One of these three simply substitutes BERT for BOW, leaving the other portions of the analysis fixed. The other two approaches attempt to make improvements after the text has been represented numerically, either by (a) predicting codes jointly rather than individually or (b) synthetically creating additional narratives to represent rare codes. While far from an exhaustive set of possible approaches to this problem, these four approaches provide a rough idea of what aspects of implementation are most likely to affect performance. We used two different criteria to characterize the performance of these models, a raw accuracy measure and a balanced accuracy measure that corrects for the relative rarity of the code.

In brief, our findings are as follows:

- As measured by raw accuracy, all four approaches performed identically. For common codes, all four approaches performed better than uninformed guessing. For rarer codes, however, no approach could manage accuracies significantly better than uninformed guessing. Even on the common codes, the accuracy achieved was relatively low in an absolute sense (roughly 75 percent).
- After correcting for code rarity, as measured by balanced accuracy, the three approaches based on BERT performed equivalently and significantly better than uninformed guessing, regardless of code rarity. The simpler BOW approach, on the other hand, performed worse (and no better than uninformed guessing) on rare codes.
- Taken together, these results indicate that the three approaches based on BERT could be used with relatively modest investment to construct a system to support human reviewers by identifying narratives likely to contain prioritized content across a wide range of rarity. Such deployments could serve to increase the effective capacity of human review. However, efforts comparable in scale to ours are unlikely to be able to produce fully automated review systems.
- The clear balanced accuracy improvement on rare codes provided by the three approaches based on BERT over the one based on BOW provides another demonstration of the value of sophisticated language models such as BERT. Relatively modest investments in more sophisticated uses of modern language models or language models trained specifically on health care data should be expected to result in further performance increases.
- As measured by balanced accuracy, performance differences among the three approaches based on BERT do not show a clear trend as a function of the rarity of the code. Rather, the relatively modest differences in performance seem to be due largely to idiosyncrasies in each code. Of note is that the approaches we used did not seem to have trouble with the concept of actionability or its components, even though it is one of the more complex codes

for human coders to capture. In contrast, a seemingly more straightforward concept—perceived technical competence—was particularly challenging for the models to fit.

Our results suggest several policy implications:

- *The success of the fairly simple models used in this pilot study supports the promise of these approaches for analyzing patient narratives at larger scale*. Overall, the models we trained to predict the human codes in narratives generally performed better than chance in our experiments. This is encouraging, but whether these types of NLP algorithms have a real role to play in efficiently scaling analyses of narrative health experience survey data also depends upon how these results play out relative to (a) different use cases and performance criteria and (b) the level of investment (in time, expertise, computing resources) that is available or required within a specific context.
- *There is labor-saving potential in leveraging the strengths of both machine and human coders, potentially in creative ways*. For example, machines could first be used to optimize the sample of narratives supplied to human coders, such that rare but important content is oversampled.
- *Coding performance was significant even with off-the-shelf computing equipment and routines and would likely improve with even modest computing investments*.
- *Perhaps the most obvious opportunity for additional investment is increasing the size of the data set on which to train the models, which we expect would improve performance*. As it stands, the limiting factor in the number of human-coded narratives available to train a model is the time and expense of human coders. Indeed, this is the very reason to explore NLP approaches and to undertake the analysis presented here. It should also not be ignored that ongoing human coding will be necessary to maintain model validity and optimization over time.
- *Efficiency may be gained by contracting model building to specialized companies*. Health care providers seeking to use NLP to scale-up their use of patient narratives can contract out the building of a model to a range of private companies specializing in precisely this task. These companies have several advantages over in-house efforts, including their ability to spread out fixed costs across multiple clients. Even when contracting out this analysis, however, in order to get value out of the resulting model, health care providers must think critically about what exactly they intend to use the model for and what this use case implies about the costs of different types of errors. Ideally, vendors should work with clients to precisely specify the use case of the model and provide statistically robust and continuously evaluated measures of their performance within that specific use case.
- *Broad stakeholder discussions could help coordinate use of NLP for patient narratives*. Health care providers and organizations tasked with improving the quality of health care could also consider organizing ongoing broader discussions among stakeholders, including patients, about the role of NLP in analyzing health care data. Once priorities are identified, a public benchmark that measures the ability of a model to achieve those priorities and an associated dataset could be produced and released.

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
AI	artificial intelligence
AUC	area under the curve
BERT	Bidirectional Encoder Representations from Transformers
BOW	bag of words
CAHPS	Consumer Assessment of Healthcare Providers and Systems
CG-CAHPS	CAHPS Clinician & Group Survey
GLUE	General Language Understanding Evaluation
MHQP	Massachusetts Health Quality Partners
ML	machine learning
NCQA	National Committee for Quality Assurance
NIS	Narrative Item Set
NLP	natural language processing
PCMH	Patient-Centered Medical Home
SMOTE	Synthetic Minority Over-Sampling Technique
TAG	technical advisory group
TF-IDF BOW	Term Frequency-Inverse Document Frequency Bag-of-Words

1. Introduction

Narratives about patients' experiences with health care contain a wealth of information about what is important to patients and are valuable in contextualizing and improving patient experience. The deidentified example in Box 1.1, collected using standardized open-ended questions that appear on systematically sampled patient experience surveys, illustrates how narratives often convey ample actionable information that (a) elaborates on existing domains such as care coordination, which are covered by closed-ended questions contained in patient experience surveys, and (b) incorporates additional domains, such as emotional rapport, which is valued by patients (Greaves et al., 2013; Huppertz & Smith, 2014; Schlesinger et al., 2020).

Box 1.1. Illustrative Patient Narrative

Dr. X is an outstanding physician. I feel very fortunate to be one of her patients. She is very thoughtful about my complex medical history and always takes very seriously any concerns that I express about my current health. She is extremely warm and caring and is also extremely smart. A very talented doctor! Once, I was experiencing an unusual symptom and Dr. X not only provided a thorough exam and ordered appropriate tests but she wouldn't let me leave the office until she had the test results. She went to great lengths to get those results because she knew I was quite concerned about what I was experiencing. It made me feel truly taken care of—I felt a level of trust and safety knowing that she would always go above and beyond to take care of me. In my experience, it is very unusual to find that level of personal commitment. I was already a huge fan of Dr. X, but that was an endearing experience that solidified my loyalty to her. Dr. X is very easy to talk. She has a good sense of humor and is very curious. I always feel well understood when we talk, and she takes my concerns very seriously. I can't say enough good things about Dr. X. She is wonderful.

The value of patient narratives for quality improvement (Grob et al., 2019; Huppertz & Otto, 2018) is matched by the challenge of organizing, analyzing, and generating actionable insights from the data, especially when faced with many narratives. Rigorous qualitative approaches to analyzing these data do exist but are labor- and time-intensive and feasible for only a modest sample of narratives (Grob et al., 2016; Lopez et al., 2012). These approaches typically involve trained researchers reading every narrative, assigning relevant codes as descriptors of meaning to segments of text, discussing and reaching agreement with other coders, and identifying important themes and insights in the data (Creswell & Poth, 2018; Kuckartz, 2014; Miles et al., 2014). Scaling such an approach to a large number of narratives might be prohibitive and impractical for health systems that collect tens of thousands of comments annually and wish to extract actionable insights from those comments in a timely and cost-effective way. Certainly, it would be impractical for monitoring performance at the level of a national health insurance program (e.g., Medicare) or most state health systems. Sampling narratives is one potential solution in the

short run, but because human coding costs build over time, there is value in building an automated processing capability.

Using Natural Language Processing to Analyze Patient Narratives

One potential solution to mitigate the resource requirements of human coding of narratives is the use of computer algorithms to extract structured meaning from unstructured natural language (Chowdhury, 2005)—that is, natural language processing (NLP; Deo, 2015; Friedman et al., 2013; Guetterman et al., 2018). NLP most frequently refers to the use of machine learning (ML) methods to analyze natural language data,[3] but not all NLP algorithms utilize ML (Cambria & White, 2014). Rules-based approaches to NLP have also been widely utilized in diverse contexts, including in predicting conditions such as pneumonia using natural language radiological reports (Elkin et al., 2008) and in extracting behaviors associated with autism spectrum disorders from electronic health records (Leroy et al., 2018). *Machine learning,* in turn, refers to a class of statistical techniques and algorithms that allow computers to "learn" patterns in data automatically.[4]

Although computer science and computational linguistics have been making advances in the realm of NLP since the 1950s, use of NLP in the realm of patient experience with health care, as measured by standardized patient surveys, is a relatively new undertaking. The most common application has been to use keyword searching and other NLP techniques to facilitate quality and safety monitoring in a number of health care settings (Divita et al., 2015; FitzHenry et al., 2013; Iyer et al., 2014; Jha, 2011; Murff et al., 2011; Rosenbloom et al., 2011; Salt et al., 2019; Wang et al., 2009). In the few studies that have reported use of NLP to analyze patient narratives, it has been used to classify comments by topic (Greaves et al., 2013), encode the sentiment expressed in comments (Huppertz & Otto, 2018), determine if comments include mention of unforeseen topics (Ranard et al., 2016), and extract themes from patient narratives already broken out by topic (Nawab et al., 2020).

Much of the value of this emerging field of NLP analysis of patient narratives is still to be realized. Several for-profit firms have made advances beyond what has been reported in the scientific literature. These firms are providing a service to health care systems that might

[3] This usage is similar to that of the term *artificial intelligence* (AI). Strictly speaking, AI refers to the use of computers to complete tasks perceived to require human-level intelligence. Thus, NLP systems are AI systems. Though modern authors and practitioners focus mostly on the use of ML for AI, some systems (e.g., expert systems) that do not use ML are still considered AI by some authors. See National Academies of Sciences, Engineering, and Medicine (2019) for an example.

[4] ML has three subdomains. In supervised ML, input and corresponding output data are supplied (in this case, from human coding), and the aim of the analysis is to map predictive functions from the input to the output. In unsupervised ML, only input data are supplied, and the aim is to model the structure and distribution of those data. In reinforcement learning, rewards are supplied to the algorithm based on its selection of one of a number of actions, and the algorithm seeks to maximize its reward. This project focuses only on supervised ML.

otherwise set patient comments aside or only scratch the surface in terms of the insights they are able to extract. The desire to protect intellectual property precludes these firms from being fully transparent about their analytic methods. As a result, NLP can appear to be a "black box," undercutting scientific progress and credibility with stakeholders (Cabitza et al., 2017; Castelvecchi, 2010; Dias et al., 2019). The inability to validate the output of such an analysis raises important concerns in the high-stakes environment of clinical practice.

Gathering Expert Input on Use of Natural Language Processing with Patient Narratives

To address these issues, on October 29, 2019, the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) research team convened a technical advisory group (TAG) to develop principles and considerations for automated analysis of patients' narratives about their experiences with health care. The TAG was asked to give particular consideration to how such an analysis should proceed when analyzing patient narratives that have been collected as part of a CAHPS survey administration.

Members of the TAG brought diverse experience to the discussion, including expertise on CAHPS surveys, the CAHPS Patient Narrative Item Set (NIS) that is a supplemental item set for the CAHPS Clinician & Group (CG-CAHPS) Survey, NLP, and health care quality improvement. We did a search to identify firms that use NLP methods to analyze patient narratives to assist health care providers and systems in identifying actionable insights in the data. Representatives from these firms were also invited to serve as members of the TAG to provide industry perspectives based on the consortium's current available knowledge of the evolving market of vendors specializing in NLP methods used in health care.

The key principles and considerations identified by the TAG were built on several assumptions, including that (a) an NLP method for analyzing patient comments needs to scale to many thousands of cases; (b) the primary analytic engine for such an analysis will be computerized, but the basic NLP analysis will be augmented by human input (e.g., construction of training datasets, rules, and ontologies) and quality control; and (c) computerized capabilities will evolve over time, and hence guidelines for conducting such an analysis will need to be periodically updated.

The TAG decided that NLP analyses of patient comments should ideally

- generate useful and actionable information, including key themes and sentiment expressed in the analyzed narratives and specific events and practices that can be targeted for improvement
- provide tools for contextualizing and organizing patient narratives, such as providing the frequency distribution of themes as context for individual narratives and the ability to link open-ended feedback generated in response to the Patient NIS to closed-ended CAHPS survey responses

- facilitate novel inference by, for example, providing insights into patient experience beyond those that come from closed-ended CAHPS measures and revealing root causes of important issues, experiences, and actionable events
- address unique challenges and opportunities in the data; for example, by providing an automated, systematic means of accounting for repetition and alternative phrasing used by respondents.

The CAHPS research team conducted this demonstration study with these considerations in mind.

This Study

Our specific overarching objective in conducting this study was to provide an open-access demonstration of the feasibility of extracting key information from a large set of narratives about patients' experiences with ambulatory care using NLP and to highlight key challenges faced when doing so. In our study, we used various NLP and ML methods to recognize the presence or absence of a specified list of concepts in the narratives. More specifically, we used a set of methods defined in the next section to predict the codes assigned to patient narratives by a team of human coders (Martino et al., 2017). In technical terms, we trained shallow fully connected feedforward neural networks on numerical representations of patient narratives constructed with (1) the term frequency-inverse document frequency bag-of-words (TF-IDF BOW) approach (see, e.g., Zhang et al., 2015) and (2) the output of the BERT (for "Bidirectional Encoder Representations from Transformers") pretrained language model (Devlin et al., 2019). We explored predicting codes both jointly and individually, as well as considering a technique designed to improve performance with rare classes of narrative.[5] We focused on the scoring and organizing potential of NLP because we believe that it is crucial to provide access to verbatim narratives, tagged with and organized by content descriptors, along with useful summary information that the NLP analysis affords.

We anticipated some challenges in conducting this analysis. First, we expected difficulty in accurately classifying patient narratives due to limited sample size (i.e., the number of human-coded narratives available). For context, in recent work comparing the efficacy of a few different types of models for supervised classification of full-text data across several different datasets (Zhang et al., 2015), the smallest dataset used contained 30,000 samples per class, analogous to narratives both with and without each code in our case. Using a model trained on the simple bag-of-words (BOW) approach—one of the techniques used here—they observe an accuracy of just under 90 percent. Their task, assigning the category of a news article (e.g., business, politics,

[5] As described more fully in Chapter 2, the Synthetic Minority Over-Sampling Technique (SMOTE; Chawla et al., 2002) seeks to improve performance on rare codes by creating new synthetic narratives that are similar but not identical to true minority class narratives.

or science and technology), was arguably easier than coding patient narratives, and they used much more input data. Having a limited sample size is likely to be especially problematic for very rare classes,[6] where the model may have as few as 50–100 human-coded narratives from which to learn. Second, we anticipated that it might be difficult to identify actionable insights contained in narratives (i.e., sufficiently detailed information about an experience that can plausibly be used to modify problematic health care practices and emphasize effective ones), as actionability is a more complex and challenging aspect to identify even among human coders (Grob et al., 2019). Regardless of whether these are solvable issues, expert human judgment will still be critical, at some level, to ensure that important narrative elements are not overlooked.

In Chapter 2, we provide a brief and basic primer on the techniques used in this exploration, organized around the two steps that a computerized system designed to classify narratives (e.g., into codes like those that our human coders apply) must perform. These include representing natural language data numerically in some vector space and then classifying the data based on that representation. In Chapter 3, we describe the data analyzed in this study. In Chapter 4, we provide a set of stylized results to illustrate the potential for and challenges of applying NLP to such data. In Chapter 5, we summarize our findings, draw conclusions, and discuss the limitations of our study.

[6] The methodological field of human qualitative analysis tends to describe tagging text with *codes*, whereas the methodological field of ML tends to refer to *classes*, which in this case can be the presence or absence of a code. Here, we use the terms *code* and *class* interchangeably.

2. Background on Machine Learning Approaches to Natural Language Processing

To apply ML methods to classify narratives, we must first represent each narrative passage numerically (more formally, as a vector in a high dimensional space) so that the data can be used by a computerized ML algorithm to predict whether each narrative belongs to a class in which we are interested (e.g., whether the narrative pertains to provider communication). The goal of such an analysis is to create a way to map the complex and diverse features of human language onto a judgment of whether or not a specific narrative belongs to a class.[7] The effort builds off of a set of narratives that have already been classified by humans. These known "gold standard" classifications are used to train the computer model to recognize narrative features that map onto each class of interest (e.g., the machine identifies that the word *explain* is most common in narratives pertaining to provider communication).[8] Once trained, this model is then used to predict the classification of previously unclassified narratives.

In contrast to ML, an alternate approach to building a computer system to automatically code these narratives would require encoding expert knowledge in explicit and predefined rules. This is difficult because of the challenge of identifying the massive number of potential narrative features and mapping them explicitly onto the presence or absence of each of our codes of interest, based exclusively on expert knowledge. ML circumvents this task by considering instead a family of functions to represent some partial knowledge about the structure of the problem and select which specific operationalization gives the "best" performance on human-coded data. Compared with regression analysis, which aims to define an optimal linear model (i.e., the line that best fits the data from the set of all possible lines), ML algorithms often rely on considerably more complex models and on techniques that iteratively improve the fit and converge on a reasonable solution.

For context, it is helpful to consider how this might operate in a domain other than NLP. For example, take the case of image classification, where the color of each pixel in an image can be naturally represented numerically and packaged into a set of features (i.e., a vector) with little effort. One can imagine assembling some number of images, some of which contain, for example, a panda, and some of which do not. Having labeled these images by hand, we can then pass our dataset to an ML algorithm to try to build a model that predicts the presence or absence

[7] More formally, the goal is to identify a function that translates all the points in the high-dimensional space of all possible narratives into a binary variable that indicates whether a specific narrative belongs to a class.

[8] The use of quotations is meant to convey that this standard is less than perfect, as the capacity of humans to code depends on skills, training, and experience. Ideally, this part of the process should be done as transparently as possible so that consumers of the results of an ML analysis can judge how optimally the analysis has performed.

of a panda in any image based only on an analysis of the numeric pixel values that constitute the image. In the case of patient narratives, however, the immediate problem is that we do not have numerical features for analysis but strings of text. To take advantage of the rich library of ML techniques, we must first map strings of text to numerical vectors—the representations described in the section "Machine Learning Using Text."

Universal Challenges in Machine Learning

Consideration of familiar statistical techniques, such as least squares and logistic regression, can help anticipate potential pitfalls when using ML approaches.[9] When one is designing a model, increasing complexity (e.g., adding new features or predictor variables) can improve fit because the model is able to use more information to make its predictions. But as one continues to add more complexity to a model, one eventually reaches a point where one stops predicting generalizable features and instead starts capturing random idiosyncrasies of the specific data used to train the model. This is the problem of *overfitting*.

In the case of classical regression analysis, there are well-developed techniques to address overfitting and goodness of fit. More complicated models, such as those used here, require a different standard of model evaluation (Breiman, 2001). Roughly speaking, unlike the simple list of predictors in linear models, the complexity of models generated by ML precludes meaningful simple model interpretation. Because of this, we need another means of evaluation, and hence the focus shifts to the model's ability to predict reserved test data it did not encounter in training (i.e., its ability to generalize accurately from training data to test data). The performance on the test data is likely to be worse than the performance on the training data, since these models are large and complex enough to learn idiosyncratic relationships specific to the training data in addition to more general relationships that will serve as robust predictors. Put another way, an overfit model will "memorize" the training data but perform worse on unseen test data.[10]

A simple random splitting of the available human-coded data does not fully utilize all of the available data. The test set is unavailable to train the model, and the training set cannot be used to provide statistical power for the model evaluation. One solution to this, which is used here, is n-fold cross-validation. Essentially, instead of a single partitioning of the data (e.g., nine-tenths for training data, one-tenth for test data), the data are partitioned into multiple subsets (e.g., ten one-tenth subsets), and a model is fit multiple times with each of the partitions taking a turn as

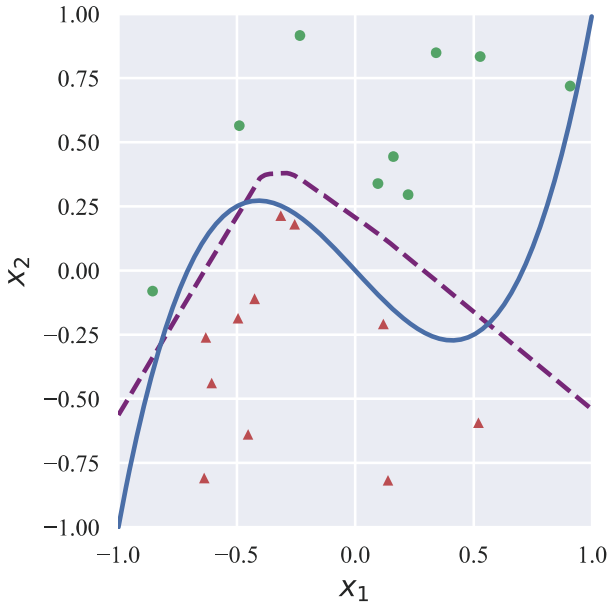
[9] Under certain circumstances, these statistical techniques are special cases of the ML approaches used here.

[10] A related problem can arise when random subsamples are of insufficient size to ensure representation of the population. In this case, two such subsamples could each capture slightly different generalizable features, and a model designed to fit one may not fit the other as well. For purposes of the exercises here, we set this consideration aside and focus instead on the more persistent problem of overfitting.

the test data. These multiple fittings (i.e., performance in accurately predicting the test data) are then compared and summarized.

The second pitfall derives from having *small data sets*—that is, those that are insufficiently sized for the problem at hand. Having few data points makes it more likely that the specific data set does not adequately represent the population of interest. This can manifest as "holes" in the data where the model does not have any nearby data points to fit. If these gaps fall along boundaries between classes, this can make it harder for the model to accurately draw those boundaries. This problem is akin to having too little power in traditional inferential statistics Figure 2.1 presents a graphic example, where a classifier tries to find the true boundary between two classes. Cases with a given code (circles) and without that code (triangles) are represented in a two-dimensional space (in practice the machine is working in many dimensions). The sparse data in Figure 2.1 caused the classifier-estimated boundary (dashed line) to miss that the true class boundary (solid line) has changed directions.

Figure 2.1. Illustration of the Challenges of Insufficient Data



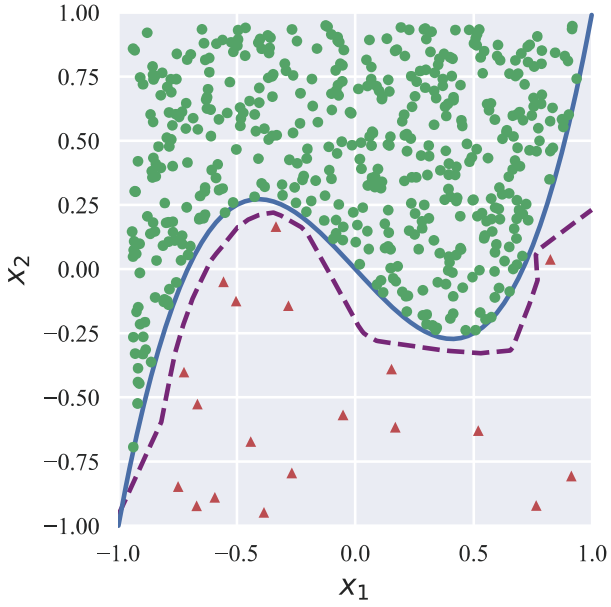
NOTE: The true class boundary is in blue, cases with a given code are denoted by green circles, and those without that code are indicated by red triangles. The purple dashed line is the boundary learned by the classifier.

Challenges Particular to Rare Events

Reticence to misclassify rare events might stem from dire real-world consequences specific to the use of the algorithm-based classification. For example, medical errors are likely to be reported by only 3 to 5 percent of ambulatory care patients but are so consequential for quality

that identifying them is of paramount importance. In our case, the differential impact arises in cases where we intend to deploy these algorithms to find rare "needles" in a large "haystack" of patient narratives. For example, because patients rarely make negative statements about provider technical competence,[11] the presence of this code (one class) is much less frequent than absence of this code (the converse class). In the ML literature (Elkan, 2001)—and in screening in general (De Smet et al., 2004)—this is typically modeled through different "costs" for misclassifying each of the two classes. These costs, combined with the relative prevalence of the two classes, determine whether it is less preferable for the model to err on the side of guessing positives or to err on the side of guessing negatives. Together, these conditions are called the *operating condition* of the model (Flach, 2003).

Figure 2.2. Illustration of the Challenges of Uneven Data



NOTE: The true class boundary is in blue, cases with a given code are denoted by green circles, and those without that code are indicated by red triangles. The purple dashed line is the boundary learned by the classifier.

In general, when dealing with classification problems, predictive power is greatest when classes are relatively evenly distributed. The situation where one class is more prevalent than the other is referred to as *class imbalance*. The more common class in these situations is referred to as the *majority class* and the less common class as the *minority class*. When imbalance is pronounced, as in the case of negative statements about technical competence, it is difficult to

[11] For a complete description of the codes assigned to patient narratives, see Chapter 3.

train a model because there are fewer data points within the rare class to provide information on the precise location of class boundaries. Figure 2.2 presents a simple example of this case, where the presence of the class (circles) is much more prominent than its absence (triangles). Despite a large volume of data, the paucity of absent cases has resulted in the classifier mistaking the location of the boundary in the middle of the right-hand side.

We anticipate this pitfall with rarely mentioned but important aspects of care. A number of techniques exist to attempt to mitigate this challenge. Generally speaking, these techniques involve assigning additional importance (technically, weights) to minority class datapoints during training or modifying the training data in some way to render it less imbalanced. One technique for modifying the training data, *undersampling*, involves discarding majority class datapoints. Another, *oversampling*, involves adding members of the minority class, most commonly by simply repeating datapoints already in the training data.[12]

Because rare codes such as negative statements about care coordination are often important in an analysis of patient narratives, we explored the use of one such technique, the Synthetic Minority Oversampling Technique (SMOTE) as a means of improving model performance. SMOTE attempts to address class imbalance by introducing synthetic, or newly created, minority class data points rather than other techniques such as oversampling existing data points (Chawla et al., 2002).[13] One can conceive of SMOTE as attempting to add examples to the rough area where true minority class examples have been found, rather than adding them precisely where they have been found. A comparison of such techniques is outside the scope of this work, and our exploration of SMOTE is primarily intended to illustrate the potential of such approaches. SMOTE was specifically chosen over undersampling to make full use of our available data and over simple oversampling due to the tendency of that technique to produce less general classifiers (Chawla et al., 2002).

We also addressed the challenge of rare codes by considering two different measures of model performance. Our first and simplest metric was *accuracy*, or the percentage of narratives in the test set correctly classified by the model. The second metric was *balanced accuracy*, or the average of the percentage of positive examples classified correctly (i.e., the true positive rate) and the percentage of negative examples classified correctly (i.e., the true negative rate). Accuracy treats misclassification of the classes equally, which may undervalue misclassification of rare classes. Balanced accuracy, in contrast, penalizes the model more for missing rarer classes. As discussed in greater detail in Chapter 3, to the extent that rare mentions are particularly important, it would argue for paying particular attention to balanced accuracy as a metric of model performance.

---

[12] See Chawla et al. (2002) for a more thorough description of the breadth of the field and Gu and Leroy (2020) for an example of one of these techniques in a health care context.

[13] These points are constructed by randomly sampling points along the line between each real minority class data point and its nearest within-class neighbor.

Machine Learning Using Text

Since ML techniques are designed to detect and model statistical relationships between sets of numbers, if we wish to apply ML to narrative data we must convert narratives to an input format useable by ML algorithms, which is most commonly a numerical representation.

Bag-of-Words Representation

One way to accomplish this is through the BOW approach. In this scheme, each member of the data set (the string of text making up the narrative) is mapped to a vector whose entries represent the prevalence of each word in that string of text. For example, a simple scheme for representing the sentence "An orange is orange" would have a "1" in the places in the vector representing the words "an" and "is," a "2" in the place representing "orange," and zeros in all other places in the vector (since no other words appear in this phrase).[14] This simple-to-implement technique can be effective in some situations. For example, spam filters frequently use this representation to recognize unsolicited and unwanted messages (Alkahtani et al., 2011). However, BOW provides no information on structure present in language, such as terms often appearing together or synonyms, since it relies solely on word frequency.

Learned Text Embeddings

One way of including semantic information is to use pretrained embeddings that represent some segment of text (either an individual word or a longer passage) numerically as a vector. The representation of this segment is derived by optimizing it for some general prediction task or tasks for which a very large dataset can be assembled (such as predicting the missing word in a sentence). This representation can then be used for other prediction tasks, on the theory that much of the information retained by the embedding will be general information about the structure of language rather than information specific to the pretraining task(s). A number of options for these embeddings exist, including word2vec (Mikolov et al., 2013) for word-level embeddings and the Universal Sentence Encoder (Cer et al., 2018), which is optimized for sentence-length segments of text. In our study we utilized BERT (Devlin et al., 2019) to generate text embeddings, since our work commenced shortly after this model achieved a marked improvement in state-of-the-art performance as measured by the General Language Understanding Evaluation (GLUE) benchmark. However, at the time of writing, rapid progress continues to be made on machine learning approaches to NLP, and BERT is no longer the state of the art.[15]

---

[14] This example is provided for illustrative purposes only. This work actually uses TF-IDF for our BOW representation. More detail is given in Chapter 3.

[15] The current leader as measured by the GLUE benchmark is StructBERT (Wang, 2019).

While one can certainly utilize BERT to generate embeddings in this way, this is not the way in which BERT or its descendants, like StructBERT, were used to perform the classification tasks that generated their GLUE scores. Instead, the models were fine-tuned for each specific classification task. This process uses the BERT-like model not as a static embedding but as a building block of a larger model.[16] Put another way, fine-tuning allows the embedding to continue to adapt using task specific knowledge even after pretraining is complete. As the BERT model is quite large, we did not have the computational resources to use this technique. We expected, however, that it would likely improve the power of the models trained.

[16] Technically, the pretrained model is used as the first layer of a deep neural network whose output layer is adapted to the task at hand. The weights of the layers derived from the pretrained model are initialized to their values at the end of pretraining and any additional layers are initialized as normal. The whole model is then trained on the data of interest.

3. Data Source and Methods

This chapter provides an overview of our modeling approach, describes the source of the narrative data, and summarizes the human coding of those data. It ends with technical detail on ML methods, which will be of interest primarily to technical audiences familiar with ML.

Overview of Our Modeling Approach

We employed four approaches to automatically code patient narratives. Across the four approaches, we held the type of classifier constant while varying inputs and resampling techniques. As a baseline, the training input for our classifiers was a simple BOW representation of each narrative. To explore the relative utility of the more complex language representation afforded by modern language models, we compared this first approach with three that utilized the open-source language model BERT to provide a numerical representation of each narrative as an input to the classifiers. The first of these three BERT-based approaches was otherwise identical to the BOW approach, and hence the comparison of these two approaches illustrates the impact of a more sophisticated language representation. We refer to this second approach as *BERT*. The remaining two approaches attempted to make improvements downstream of the language representation. One, which we refer to as the *joint model*, slightly modified the design of the classifier to allow it to share information between the different codes, forcing it to simultaneously predict all 26 jointly rather than each independently (as is done in both BOW and BERT approaches). Finally, we explored pre-processing the training data with SMOTE after passing it through BERT (referred to as *SMOTE* below). These approaches were evaluated using ten-fold cross validation to make maximal use of our available data.

To measure performance of these approaches, we considered two different measures of the power of our models, representing two broad use cases. Our first and simplest metric was *accuracy*, the percentage of narratives in the test set correctly classified by the model. That is,

$$Accuracy = \frac{Number\ of\ Narratives\ Classified\ Correctly}{Total\ Number\ of\ Narratives}.$$

This metric corresponds to an operating condition in which the costs of the two types of errors are identical. As a baseline against which to compare the performance of the model, we report the *base rate* for each code: the prevalence of the majority class as estimated by the data. This represents the performance of a simple classifier that always guesses the majority class.

The second metric was *balanced accuracy*, the average of the percentage of positive classifications that were correctly applied (i.e., the true positive rate) and the percentage of negative classifications that were correctly applied (i.e., the true negative rate). That is,

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{Correct\ Positive\ Classifications}{All\ Positive\ Narratives} + \frac{Correct\ Negative\ Classifications}{All\ Negative\ Narratives}\right).$$

This represents an operating condition in which the cost of misclassification increases with the rarity of the code missed. Guessing randomly and guessing the majority class both have the same balanced accuracy: 50 percent. To the extent that rare mentions are particularly insightful, this circumstance would argue for paying particular attention to balanced accuracy as a metric of model performance.

Human-Coded Data Source

Data for this study came from the Massachusetts Health Quality Partners (MHQP) Patient Experience Survey. This statewide survey, conducted annually in Massachusetts since 2005, asks patients to report their experiences with a specific primary care provider and with that provider's practice. The 2017 survey was based on the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Patient Centered Medical Home (PCMH) Survey developed by researchers from the National Committee for Quality Assurance (NCQA) and the CAHPS Consortium (Hays et al., 2014; Scholle et al., 2012).[17] There were two survey versions, one focused on the experiences of adult patients and the other focused on the experiences of pediatric patients. Our study focused only on responses to the adult version of the survey.

Solo and dual practice sites were included in the survey only if they or their provider organization opted to fund the sampling of their patients. Eligible adult patients (18 years and older) had to have made a visit to a primary care provider during the previous year (i.e., the assessment year) and to have been enrolled in one of the participating commercial health plans operating in Massachusetts. Practices were required to have at least three eligible primary care providers of the same specialty, each having a panel size of at least 20 eligible patients. The 2017 survey was administered in English by mail with an option to complete online and was sent to over 205,000 adult patients who received care from 866 primary care practices statewide. The response rate to the survey was 20 percent, comparable to response rates achieved in other regional health care survey efforts (Fowler et al., 2019).

The survey contained 56 closed-ended questions followed by the five-question CAHPS NIS. The NIS, which is described in detail in Box 3.1, has a logical flow that helps a respondent to build a complete, balanced, and meaningful narrative (Schlesinger et al., 2020). The first question establishes what the patient looks for in a health care provider, thus providing an important contextual cue for answering the subsequent questions. The second question asks how the focal provider measures up to the patient's standards. The third and fourth questions ask, respectively, for descriptions of specific positive and negative experiences. The final question

[17] The MHQP Patient Experience Survey includes all questions from the CAHPS PCMH Survey plus a small set of additional questions.

elicits specific information about how the patient and provider relate to one another. Together, these five open-ended questions are designed to generate a fulsome narrative that is useful to health care providers, staff, administrators, and consumers.

Approximately 12 percent of all adult patients who completed the 2017 survey provided a response to at least one of the five NIS questions. In all, we included 4,219 comments (one per respondent) from adult patients who received care from one of 780 practices. All comments were deidentified (i.e., provider and patient names, practice names, and any statement about a rare or specific condition that could potentially be used to identify a patient were removed). The remaining text averaged 78 words (response to all five questions combined; standard deviation = 78 words) with a minimum response of five words and maximum response of 797 words.

Box 3.1. Five-Question CAHPS Narrative Item Set

In your own words, please describe your experiences with this provider and his or her office staff, such as nurses and receptionists.

1. What are the most important things that you look for in a healthcare provider and the staff in his or her office?

2. When you think about the things that are most important to you, how do your provider and the staff in his or her office measure up?

3. Now we'd like to focus on anything that has gone well in your experiences in the last 6 months with your provider and the staff in his or her office. Please explain what happened, how it happened, and how it felt to you.

4. Next, we'd like to focus on any experiences in the last 6 months with your provider and the staff in his or her office that you wish had gone differently. Please explain what happened, how it happened, and how it felt to you.

5. Please describe how you and your provider relate to and interact with each other.

Two coders independently coded a total of 1,490 (745 each, randomly selected) of these 4,219 comments. The coders had extensive experience coding NIS responses. They were previously trained to use a more elaborate coding scheme than was used for this study and achieved substantial to outstanding inter-rater reliability (kappa > 0.7) on all elements. For this study, the coders judged (1) presence/absence of each of six dimensions of patient experience (doctor-patient communication, office staff, access to care, care coordination, perceived technical competence, and emotional rapport), (2) valence of commentary within each of the six domains (if any), and (3) actionability. We defined a passage as actionable if it contained sufficiently detailed information about an experience that could plausibly be used to modify problematic health care practices and emphasize effective ones (Grob et al., 2019).

Coding actionability required several judgments on the part of coders. First, the coders judged whether the response cited a concrete action, attribute, or practice on the part of the health care provider. Next, the coders judged whether the experience described was (1) negative and preventable, (2) positive and exemplary (i.e., an experience so positive as to be extraordinary), or (3) positive but not universal (i.e., an established norm of patient-centered care that is nonetheless inconsistently practiced). Third, the coders judged whether the response contained detail about the specific person involved or the role of that person (e.g., doctor, nurse, or receptionist), where the experience occurred (e.g., the exam room), or when it occurred (e.g., upon leaving the health care facility). Responses conveying negative experiences were determined to be actionable if they (1) described a concrete action, attribute, or practice and (2) provided enough detail about who, where, or when. Responses conveying positive experiences were determined to be actionable if they (1) described a concrete action, attribute, or practice that was either negative, exemplary, or desired but not universal and (2) provided enough detail about who, where, or when.

Summary of Human-Coded Data

Of the 1,490 narratives coded by the human coders, 12 (< 1 percent) were judged to have no codable content. Table 3.1 shows the percentage of the remaining 1,478 narratives that were assigned each of the 26 codes by the coders. As this table indicates, the dimensions of patient experience that were most commonly identified in the narratives were provider communication (58.3 percent) and emotional rapport (51.3 percent). The least commonly identified dimension was care coordination (17.2 percent). Positive comments about a dimension of care were far more common than negative comments. Negative comments about perceived technical competence, emotional rapport, and care coordination were identified in less than 3 percent of narratives. In all, 40.8 percent of the narratives were judged to contain actionable information. Table 3.2 provides examples of fragments of narratives judged by human coders to be positive or negative statements about the six dimensions of patient experience of interest. We have separated these fragments from the larger narrative in which they appeared so that the reader is able to see clearly the specific text that prompted the code; however, we remind readers that codes are applied to narratives as a whole and not to fragments of narratives. Below Table 3.2 is a box (Box 3.2) illustrating a narrative that was judged by human coders to contain actionable content on the basis of the specific details it provides about a negative experience.

Table 3.1. Descriptive Data on the Content of Narratives (*N* = 1,490) Coded by Human Coders

Dimension of Care	Any Mention %	Any Positive %	Any Negative %
Provider communication	58.3	54.8	5.5
Emotional rapport	51.3	48.5	2.9
Access to care	31.2	24.2	9.0
Perceived technical competence	28.1	26.2	2.1
Office staff	24.7	19.7	6.1
Care coordination	17.2	13.4	4.8
Actionability	**% Yes**		
Concrete action, attribute, or practice mentioned	71.5		
Negative	18.2		
Exemplary	3.9		
Notable but not universal	19.1		
Sufficient detail about “who”	35.8		
Sufficient detail about “where”	29.3		
Sufficient detail about “when”	31.0		
Actionable	40.8		

Table 3.2. Examples of Statements Prompting Certain Codes

Code	Narrative Fragment
Provider communication, positive	[My doctor and I] talk openly and comfortably. I feel respected and listened to.
Provider communication, negative	Doctor X did not make eye contact with me once during my office visit, and it was the first time we had met. It was a check-up appointment and he spent about 5 minutes with me. He did NOT ask me if I had any other questions, and quickly left the room. I was completely stunned because I DID have other things I wanted to bring up.
Emotional rapport, positive	Doctor X always makes me feel at ease to tell her what's wrong. She . . . is reassuring when you are worried. I always feel like she is genuinely concerned about my health and well being.
Emotional rapport, negative	There is absolutely no personal connection between Doctor X and me.
Access to care, positive	I was not feeling well and I got an appointment right away.
Access to care, negative	In all honesty, I have not seen my primary care doctor in years. Mainly because appointments with him would often mean waiting in the waiting room for long periods of time (2–3 hours).
Perceived technical competence, positive	[My doctor] is knowledgeable about my conditions and how I can improve them.
Perceived technical competence, negative	Doctor X is so ignorant of the very basic medical knowledge that I cannot believe she works in a prestigious hospital in this area.
Office staff, positive	When at the front desk before and after an appointment, they always make you feel like you are the only person that matters at that time.
Office staff, negative	The administrative staff . . . are very unfriendly and create an atmosphere that patients are interrupting them. They do not greet patients—often focusing more on other tasks than on patients who are directly in front of them.
Care coordination, positive	The drug that I was on for high cholesterol was giving me joint aches. Doctor X told me to stop taking it and we switched to another drug. She did blood tests . . . to see if the new drug was working. She responded immediately with test results and suggested that I stay on the [new] drug because it was working.
Care coordination, negative	Doctor X has been a bit slow to take up the online services provided by [my plan] or to receive treatment notes from specialists and to leave notes for them as well as me.

Box 3.2. Example of a Patient Narrative with Actionable Content

I didn't feel like my last visit wrapped up completely. I had a number of different complaints, and we left several of them hanging. I was told to tape my injured thumb and seek physical therapy for my hip pain, but wasn't given any instruction about how to do either.

Technical Details of Our Machine Learning Methods

This section gives a thorough technical description of the methods employed to build and train the classifiers. It is anticipated that this will be of most interest to those with deep familiarity with machine learning who are interested in the details of our approach.

We explored two modes of assigning features to the narratives. First, the pretrained BERT (large, uncased) model distributed with the Transformers NLP python module was put into evaluation mode and evaluated on the response to each of the five questions of the NIS individually, with a [CLS] token prepended and a [SEP] token appended. The output embeddings for each token were averaged across each question of the NIS, and the five resulting vectors were concatenated to form a single vector embedding of each patient response with a dimension five times the hidden dimension of the large BERT model, $d = 5{,}120$.

Second, we concatenated the responses to each of the five members of the NIS and removed words from the stop word list distributed with the Natural Language Toolkit. The resulting text was then fed into the Keras Tokenizer class set to TF-IDF-BOW mode. Due to the number of words in our corpus, the resulting vector had dimension $d = 5{,}030$.

Using these two representations, we trained four sets of models in Keras. Using each representation, we built a series of 26 identical neural network models with a single hidden layer and a one-dimensional output layer to predict each of the individual codes. We refer to these models as *BERT* when they are trained on the BERT embedding and *BOW* when they are trained on the BOW vectors. We also created a model with a single hidden layer and a 26-dimensional output layer trained on the BERT embedding to predict all 26 codes simultaneously, referred to as the *joint model*. Finally, we trained an additional instance of the 26 individual models on the BERT embedding with the training set preprocessed with SMOTE, as implemented in the python package imbalanced-learn, referred to simply as *SMOTE*. All models were trained using the Adam minimizer and early stopping to minimize relative entropy loss, as implemented in Keras. The activations in the hidden layer were rectified linear units, and the activations in the output layer were sigmoid. Note that this extends the activations of the output layer of the joint model, as the codes are not mutually exclusive.

We varied the learning rate (10^{-3}, 5×10^{-4}, and 10^{-4}), early stopping patience (1, 2, and 5), hidden layer size (512 and 1,024), and batch size (10, 15, and 25) as hyperparameters. Eighty percent of the data was split 20 percent–8 percent–72 percent for use as a validation set, early stopping set, and training set during hyperparameter tuning. Models were trained on the training

and early stopping sets for each hyperparameter configuration and evaluated using areas under the curve (AUCs) on the validation set.

Since we report accuracy and balanced accuracy, we chose a threshold to maximize each for the selected hyperparameter configuration using results on the validation set. Thus, the reported balanced accuracy results from classifying according to a different threshold than that used to calculate the reported accuracy. To accomplish this, the probability density of scores for positive and negative examples was estimated from the observed scores on the validation set using the kernel in Equation 4 of Botev et al. (2010) with the bandwidth set by leave-one-out likelihood maximization (Rudemo, 1982). The density itself was constructed using SciPy's gaussian_kde function and the likelihood maximization was carried out using SciPy's minimize_scalar. Using these estimated score densities, we calculated an expected accuracy and an expected balanced accuracy as a function of threshold by calculating expected accuracy without class skew and with class skew equal to the observed class imbalance on the validation set. The thresholds were then chosen to maximize their respective accuracies, again using minimize_scalar.

A final model was then trained on the configuration with the best AUC for each model type using ten-fold cross validation. Note that SMOTE was not applied to the test, validation, or early stopping sets. Metrics were computed by assuming that the error probabilities were independently and identically distributed across folds. That is, metrics were computed under the assumption that the inducer is stable in the terminology of Kohavi (1995). Accuracies and the confidence intervals thereof were calculated using the Wilson interval (Brown et al., 2001) with the point estimate being the maximum likelihood estimate. Note that in the case of balanced accuracy, this means that the Wilson intervals for the true positive and true negative rates were separately calculated and used to construct a confidence interval for balanced accuracy. The AUCs and confidence intervals thereof were calculated by recognizing the AUC as the Wilcoxon two-sample U-statistic, calculating the standard error of this statistic via Bamber's method, and utilizing asymptotic normality (DeLong et al., 1988). The construction of multiple simultaneous confidence intervals was accomplished via the Bonferroni correction (Casella and Berger, 2002). Statistical significance was assessed using the extent of these confidence intervals, which in the case of the accuracy metrics is equivalent to score tests with overall confidence level controlled by the Bonferroni correction.

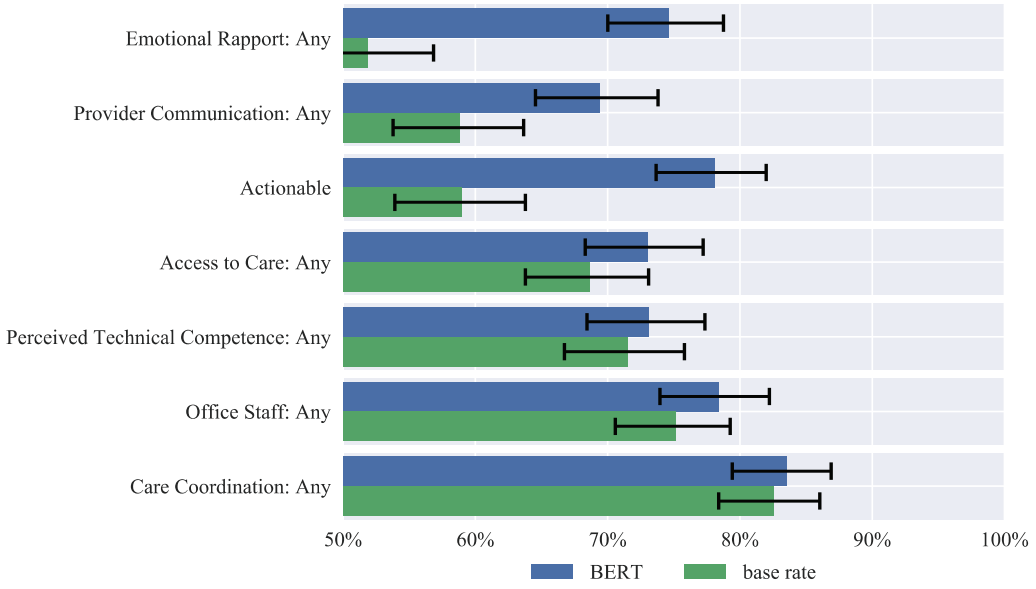
4. Performance of Demonstration Models

Performance as Measured by Accuracy

The four sets of models performed with nearly the same degree of accuracy, showing no statistically significant differences. Therefore, for clarity and brevity, we focus on results for BERT, as it represents the core approach used here. Accuracy results for all four approaches are presented in the Appendix. As described in Chapter 3, we also present the base rate, or proportion of narratives that are in the majority class, as measured using the whole dataset. The base rate represents a trivial model that simply guesses the majority class. For example, negative codes for emotional rapport were extremely rare among the coded narratives (3 percent), hence a base rate model that simply guessed that no narratives contained negative emotional rapport content would be correct 97 percent of the time. For this reason, performance relative to the base rate indicates how much value the more sophisticated model is providing, particularly on the highly imbalanced classes where bare accuracy might be misleading.

Figures 4.1 and 4.2 summarize this performance, with codes arranged top to bottom in

Figure 4.1. Accuracy Performance on Presence Versus Absence of Any Mention of Each Code, Relative to Base Rate Performance

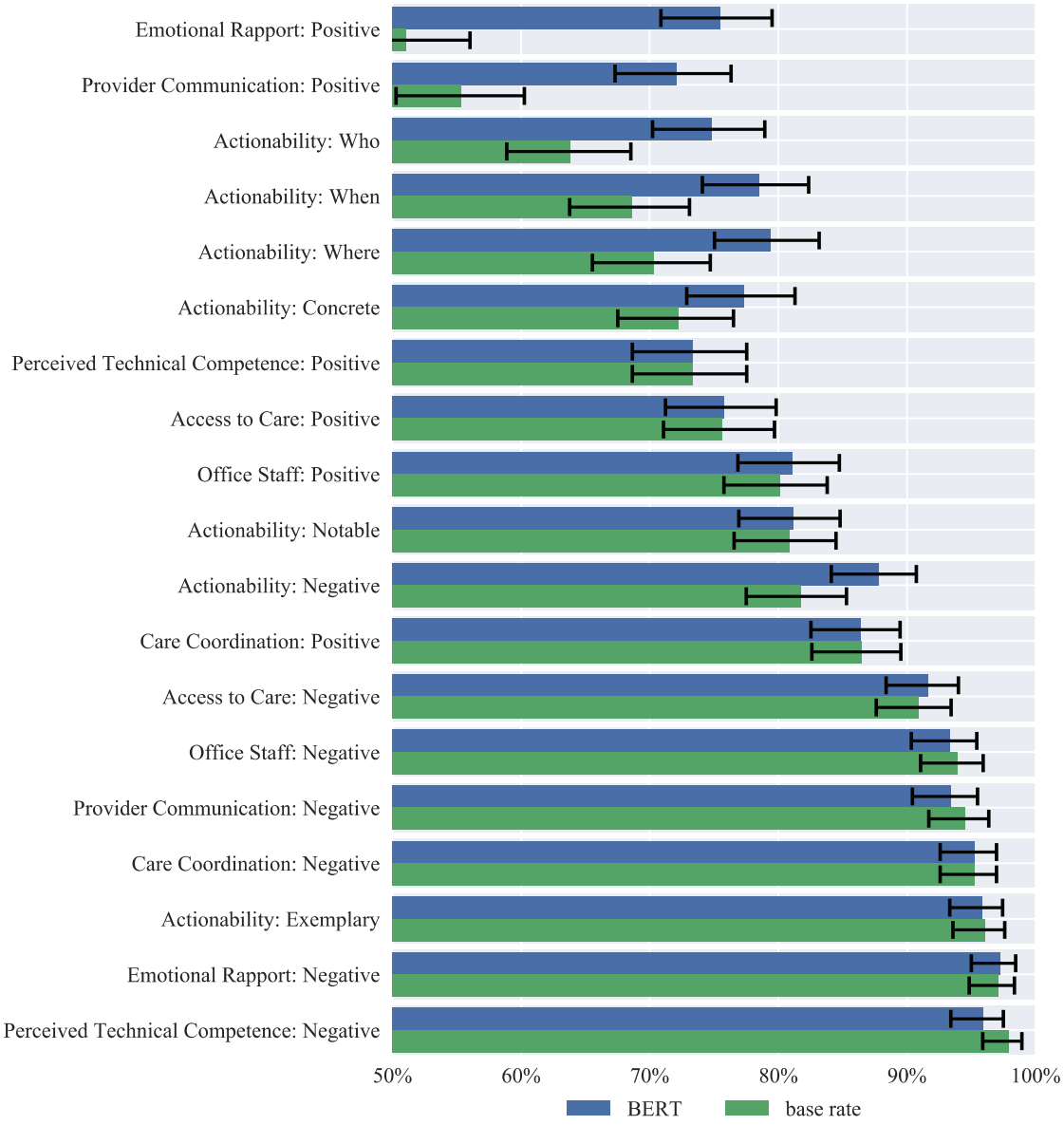


NOTE: The blue bars represent the estimated accuracy for the model, and the error bars represent 95 percent confidence intervals. Green bars give the same for the base rate. Codes are ordered from least imbalanced (top) to most imbalanced (bottom).

increasing order of imbalance. This ordering is maintained across figures. Figure 4.1 depicts results for any presence of the composite codes, plus any overall determination of actionability.

Figure 4.2 breaks out positive and negative instances, together with the components of actionability.

Figure 4.2. Accuracy Performance on Presence Versus Absence of Component Codes, Relative to Base Rate Performance



NOTE: The blue bars represent the estimated accuracy for the model, and the error bars represent 95 percent confidence intervals. Green bars give the same for the base rate. Codes are ordered from least imbalanced (top) to most imbalanced (bottom).

The model achieved accuracies in the mid-70-percent range on the more balanced codes until the base rate rose to meet this level of accuracy, whereupon the model accuracy largely tracked the base rate to within a percentage point or two. Thus, while the model provided some value on the more balanced codes (e.g., where the larger class was present in not much more than 70 percent of all data points), it provided little to no value on codes with even moderate imbalance. Moreover, where the model did provide this value, it did not provide enough to be confident deploying the model to identify narratives with the code without human supervision.

Furthermore, certain codes proved more challenging for the model to fit than others, even accounting for different degrees of imbalance. For example, despite having nearly identical base rates and performing significantly better than the base rate, the models very nearly performed significantly better (within 0.14 percent) on actionability than on provider communication (any mention). This difference is counterintuitive because the actionability code is more complex, requires greater synthesis, and is typically harder for human coders than the provider communication code.

Note that the performance of the model begins to track the base rate more closely as the base rate rises. As one might suspect, it appears that the model begins to behave like the trivial model that guesses the majority class. Though the behavior is not universal on the more imbalanced classes, several codes were assigned by the model in only a handful of cases. For example, although there were 281 narratives coded as *Actionability: Notable* by the human coders, the model identified only nine as having the code.

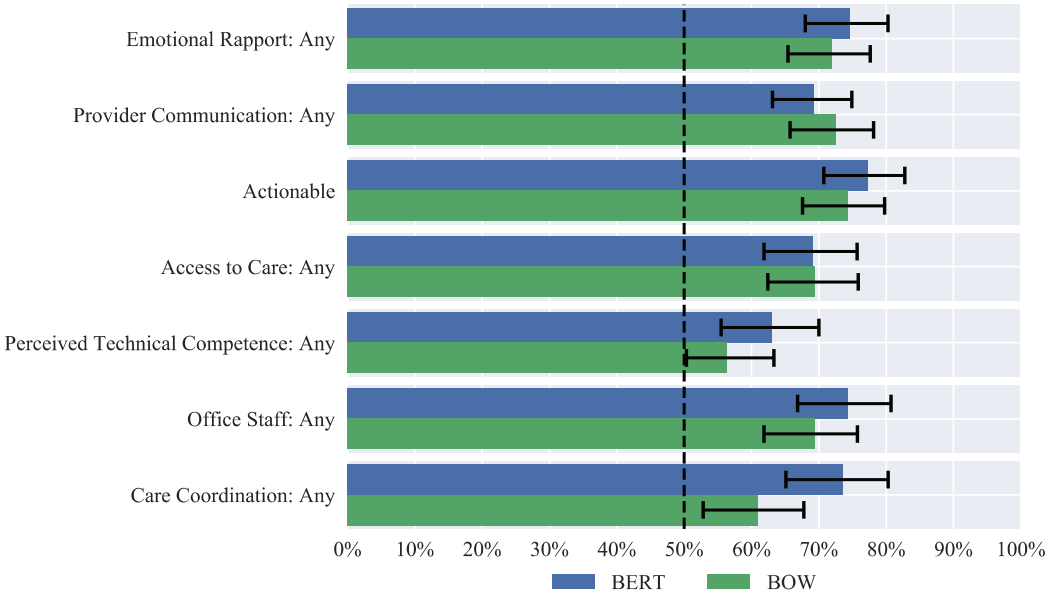
Performance as Measured by Balanced Accuracy

As discussed above, there is some evidence that the model erred on the side of guessing the majority class when judged against accuracy (as opposed to balanced accuracy). One way to probe how much the model learned about rare codes is to penalize mislabeling the minority class at the same rate as mislabeling the majority class (i.e., balanced accuracy). Recall that balanced accuracy is calculated by averaging (a) the accuracy on narratives that contain the code (i.e., the true positive rate) with (b) the accuracy on narratives that do not contain the code (i.e., the true negative rate). As such, balanced accuracy accounts for the imbalance in codes by setting the cost of misclassification proportionally to rarity. Hence, it represents a use case where identifying a rare code is especially valuable.

The three sets of models trained on the BERT embedding performed nearly identically, showing no statistically significant differences. Thus, in this section we present data from only the BERT and BOW approaches, taking the BERT approach as representative of the performance of the joint model and SMOTE. Balanced accuracy results for all four approaches are presented in the Appendix. Figures 4.3 and 4.4 are similar to Figures 4.1 and 4.2, but (a) they reflect balanced accuracy, and (b) performance is judged relative to 50 percent (dashed vertical line). We see statistically significant performance above baseline for the model trained using the

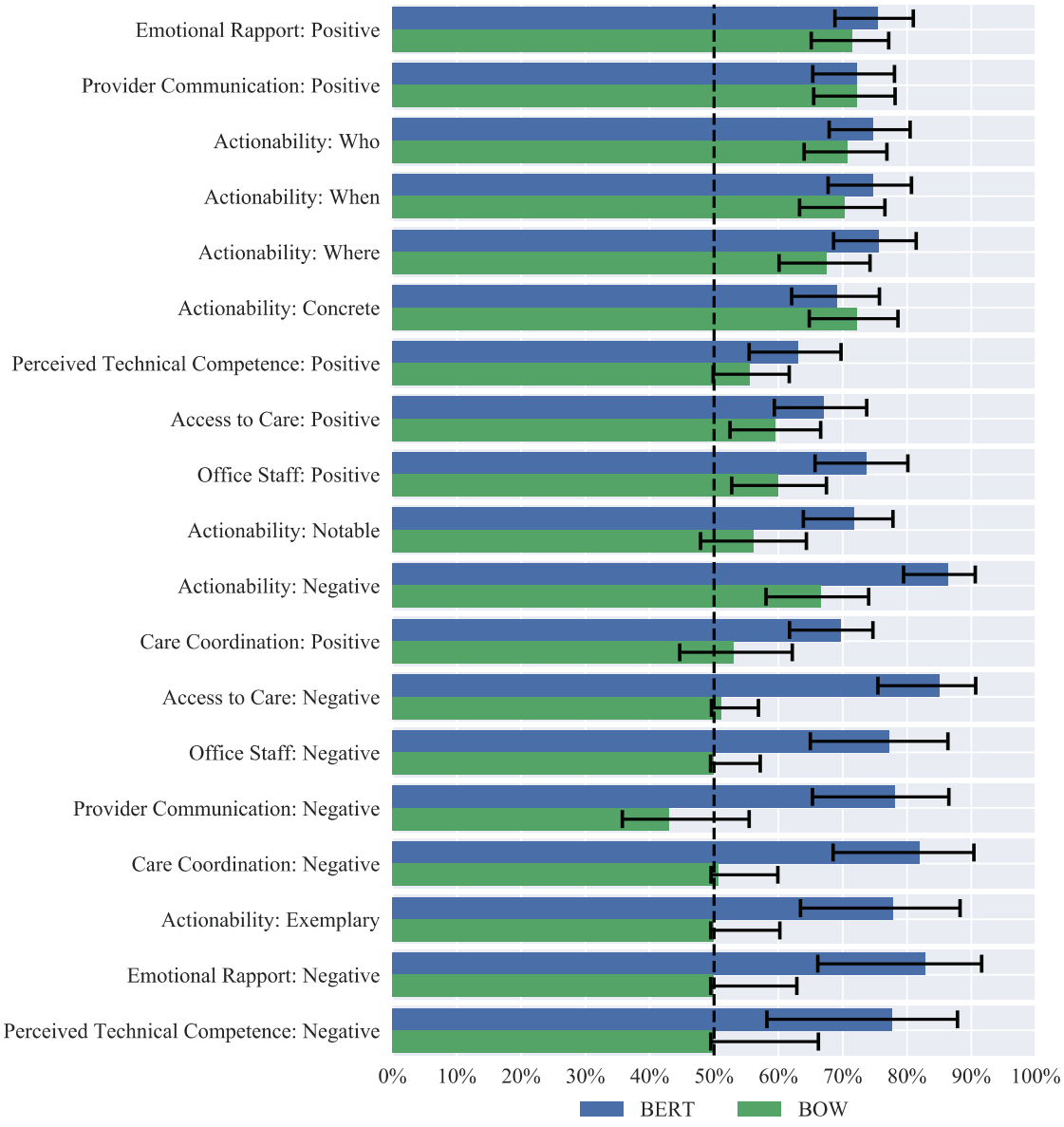
BERT embedding on all codes. On the more imbalanced codes, however, the models trained on the BOW representation showed markedly worse performance than the BERT embeddings. In one case (code *Actionability: Negative*, Figure 4.4), this simply manifests as a statistically significant performance improvement for the models trained on BERT. For all codes more imbalanced than this, however, the BOW models failed to perform better than uninformed guessing.

Figure 4.3. Balanced Accuracy Performance on Presence Versus Absence of Any Mention of Each Code, Comparing Simple BOW Approach with BERT-Embedded Model



NOTE: The dotted vertical line indicates 50 percent, which is the level of performance from simply guessing. Error bars represent 95-percent confidence intervals.

Figure 4.4. Balanced Accuracy Performance on Presence Versus Absence of Component Codes, Comparing Simple BOW Approach with BERT-Embedded Model

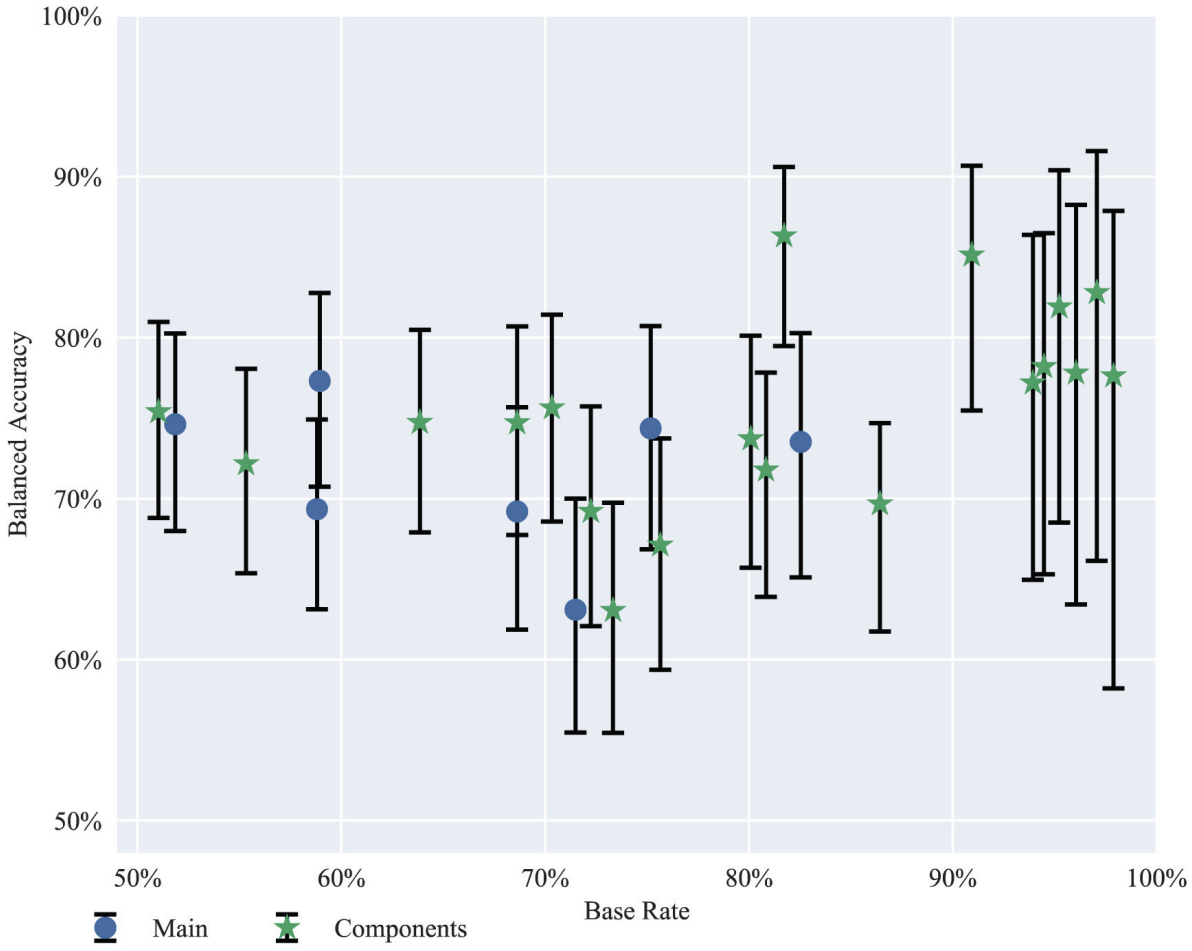


NOTE: The dotted vertical line indicates 50 percent, which is the level of performance from simply guessing. Error bars represent 95-percent confidence intervals.

Figure 4.5 explores the potential relationship between base rate and performance (as measured by balanced accuracy). Here we see little evidence of an overall trend in the balanced accuracy as a function of base rate. It appears that the predominant effect determining whether the model will perform well or poorly is the difficulty of the code itself, and which codes are most difficult seems to be consistent between the accuracy and balanced accuracy operating

condition. For example, perceived technical competence (any mention) continued to give the model considerable trouble despite having only moderate class imbalance.

Figure 4.5. Balanced Accuracy for BERT, by Base Rate



NOTE: Blue circles indicate superordinate codes, and green stars indicate subcodes. Error bars represent 95-percent confidence intervals.

The models trained on BOW failed to perform statistically significantly better than baseline on several codes for which the other three models performed adequately. This phenomenon was confined primarily to the more imbalanced codes (i.e., any code with a base rate above about 80 percent). Taken together, these results suggest that while the simple embedding performs comparably on the more balanced codes, training a model based on more realistic language features (e.g., through the pretrained BERT embeddings) is important for predicting rare codes.

5. Discussion

Summary

In this study we performed a series of experiments in fitting natural language classifiers to real-world patient narratives to explore the feasibility of using computers to automatically code narrative survey data on patient experience. We used four different approaches—three variants using a relatively sophisticated language representation (BERT) and a comparison representation based simply on word frequency (i.e., BOW). Our results provide strong evidence of the promise of such approaches, with the caveat that they would benefit from using much larger training datasets. It is notable that the training data we used are typical in size for efforts involving human coding.

Accuracy varied according to code rarity—that is, how balanced or imbalanced a code was in the data. For more balanced codes, the models showed substantial improvement over a baseline of simply guessing the more prominent class for a given code. This result itself argues for the feasibility of these methods, if provided enough data. For codes where one class was quite rare—for example, negatives statements about emotional rapport occurred in less than 3 percent of our human-coded narratives—accuracy was no better than guessing according to the base rate.

For balanced accuracy, the three BERT models performed better than chance on all codes, with no clear differences in balanced accuracy across the three BERT models. The BERT models performed demonstrably better than the simple BOW model, which failed to demonstrate any ability to identify the minority class on the codes with base rates above 80 percent. This suggests the importance of training the models using a more sophisticated language representation and provides another demonstration of the power of model pretraining in NLP. Given the benefit we observed with pretraining on general language data and the results of prior research (Gu & Leroy, 2020), it is natural to suspect that we might see further improvements with a language model pretrained on patient narratives or other natural language data from a health care context.

Finally, a particularly encouraging and intriguing result was the relatively good performance on the concept of actionability, which is much more complex for humans to judge. We anticipated that the models would also have a difficult time with this concept, so the relatively good performance of the models for actionability suggests their potential utility for more complex concepts.

Policy Implications

The models we trained to predict the human codes in narratives generally performed better than chance in our experiments. This is encouraging, but whether these sorts of NLP algorithms have a real role to play in efficiently scaling analyses of narrative health survey data also

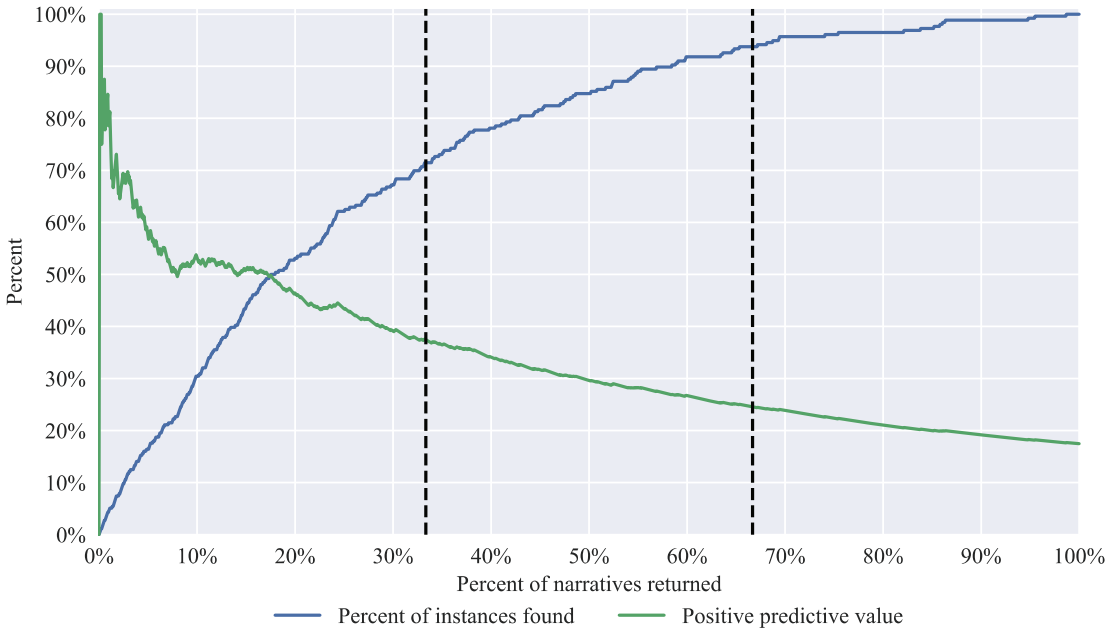
depends upon how these results play out relative to different use cases and to the level of time, expertise, and computing infrastructure users are willing to expend.

Labor-Saving Potential Lies in Combining Machine and Human Coders

The models in our experiments have not reached a level of performance that would justify deployment as stand-alone engines for analyzing patient narratives. However, even these simple models based on limited data contain useful insights about the content of patient narratives. One could imagine using them in concert with human coders. For example, consider a hospital administration that wishes to use patient narratives to understand and improve care coordination but has the resources to examine only a third of the patient narratives using human coders. If the examined narratives were selected at random, the hospital team could expect to encounter about a third of the narratives pertaining to care coordination, which could be a pretty small number. If, however, the team were to optimize a model specifically to seek out and return the one-third of all narratives most likely to contain the code *Care Coordination: Any*, they could expect to catch roughly 70 percent of the narratives that contain information pertaining to care coordination.[18] These narratives could then be subjected to human coding for greater accuracy. If instead the team could analyze two-thirds of the narratives generated, they could expect to catch essentially all the narratives pertaining to care coordination with this technique. Figure 5.1 expands this analysis to estimate what percentage of the care coordination narratives would be captured for machine-returned samples of varying sizes (blue curve). Figure 5.1 also shows how the percentage of returned narratives that actually contain the care coordination code (the so-called positive predictive value) varies by machine-returned samples of varying sizes (green curve). For example, about 17 percent of our human-coded narratives contained a care coordination code, but the machine-returned one-third sample would consist of approximately 37 percent narratives containing mentions of care coordination.

[18] Figure 5.1, the source of this estimate, was generated by allowing the threshold for positive classification to vary from 0 to 1 and plotting the resulting positive predictive value and recall against the proportion of narratives returned for each threshold. The construction is identical to that of a precision-recall curve or a receiver operating characteristic curve, with different quantities plotted.

Figure 5.1. True Positive Rate Versus Positive Rate for the Individual Models in Any Instance of the Care Coordination Code



NOTE: Vertical dashed lines occur at one-third and two-thirds. Positive predictive value (i.e., the proportion of the narratives returned that actually contain the code of interest) is given in green.

Coding Performance Would Likely Improve with Even Modest Computing Investments

As to the question of required resources, the preceding analysis was conducted using general purpose laptops not optimized for classification. This level of effort is likely to be easily replicable by any organization that employs even a few analysts with a background in statistics and computation. With modest additional investment on the scale of a few thousand dollars ($1,000 to $3,000 would likely be sufficient), such a team could proceed further by fine-tuning a state-of-the-art language model like BERT by purchasing a purpose-built machine for training models or time on a cloud computing service. Given the benefits provided even by using the BERT model for generating embeddings (i.e., learned vector representations of the text), it is reasonable to expect such an approach would result in notably improved performance. Investments in computational resources could also enable pretraining with unlabeled patient narratives or other sources of natural language health care data, either instead of or after pretraining on general language data depending on the size of the dataset that can be assembled. There is also a wide range of additional model classes and techniques for training neural networks that could be explored, given enough time for analysts to experiment with different designs.

Size of Training Datasets Is Currently a Limiting Factor

Perhaps the most obvious outlet for additional investment is increasing the size of the data set on which to train the models, which we expect would likely improve performance. Because the size of the human-coded training and test data is modest (n = 1,478), the performance of our classifiers may not fully reflect those that could be produced by a more well-resourced effort. As it stands, the limiting factor in the number of human-coded narratives available to train a model is the time and expense of human coders. Indeed, this is the very reason to explore NLP approaches and to undertake the analysis presented here.

Creatively Leveraging Both Human and Machine Coding Could Lead to Even Bigger Gains

One could imagine such an effort to accumulate more data dovetailing nicely with the care coordination case described above, especially if focused on the more imbalanced codes. So long as we are in the situation where the model is accurate enough to be a useful prioritization tool but lacks an acceptably low error rate when deployed in a standalone capacity, one could divide the human coding capacity into two efforts. One effort would involve coding narratives chosen randomly, while the other would involve coding a subset chosen by the model according to some criteria (i.e., of the remaining narratives, those likely to contain operationally important or rare codes). The random subset could be used to estimate base rates, gain insight into characteristics of interest in the patient narratives, and study the performance of the model. This could also support a deployment of the model to narratives untouched by human coders to glean what insights can be gathered with a rigorous understanding of the likely error rates. The narratives randomly selected for coding could then be fed back into the model for further training once coded by humans. The model-prioritized subset, once coded by humans, would provide more targeted insights into subjects of concern to survey sponsors. Furthermore, one could speculate that model performance might also benefit from training on the prioritized subset as well, after it has been coded by hand. The model might benefit from having an excess of rare or important codes from which to learn, due to the presence of the prioritized codes, and therefore improve its performance in a more targeted way. This process would allow the work done by human coders to compound over time, slowly improving the model as it supports their workflow.

Ongoing Human Coding Is Needed to Maintain Model Validity and Optimization

Even after the model reaches an acceptable error rate within the operating condition in which it would ultimately be deployed, ongoing validation and optimization of the model would suggest the need for further effort from human coders to generate new coded narratives. The number of coded narratives needed for this steady state is likely to be lower than for initial training, since it will be driven by the need to assemble a test set large enough to provide the statistical power necessary to measure model performance rather than the need to accumulate

sufficient data to train the model in the first place. For example, over time the base rate may shift as the hospital makes improvements in the procedures followed by its office staff (perhaps in response to insights provided by the model), leading to an uptick in the rate of positive statements about office staff present in narratives. The changing base rates will change the operating condition of the model and necessitate re-optimization. In addition, this data should be used to check for changes in the underlying statistical relationships learned by the model over time. Practically speaking, the performance of the model should be tracked to detect any decrease, which would indicate some change in the underlying process necessitating a retraining of the model.

Efficiencies May Be Gained by Contracting Model Building to Specialized Companies

Health care providers seeking to use NLP to scale-up their use of patient narratives can also obtain a completed system from a private company specializing in precisely this task. These companies have several advantages over in-house efforts, most notably their ability to spread out fixed costs across multiple clients. They can focus on just one technical area, generate a wealth of in-house technical expertise, develop modeling designs and techniques specifically for this application, and leverage access to the combined data of all their clients. However, such firms have the disadvantage of being less transparent about their methods (stemming from a need to protect intellectual property). In the long term, as collective knowledge about such analysis deepens, incentives could shift away from protecting intellectual property toward the value of implementing standard methods (as has happened with analysis of CAHPS closed-ended questions).

Even when contracting out this analysis, to get value out of the resulting model, health care providers must think critically about what exactly they intend to use the model for and what this use case implies about the costs of different types of errors. This objective defines the utility of the delivered product, more so than bare accuracy or another metric of theoretical interest. Ideally, vendors should work with clients to precisely specify the use case of the model and provide statistically robust and continuously evaluated measures of their performance within that specific use case. For example, in the deployment scenario considered above, where the model served to flag narratives for manual review, potential vendors should help clients specify a specific scenario for deployment such as whether to treat the capacity for manual review as a constraint or a cost per flagged narrative. Vendors could then use this information and further consultation with the client to develop statistically rigorous, understandable, and actionable metrics that measure performance of the deployed models relative to those goals: for example, a confidence interval for the proportion of the narratives of interest that will be detected given actual client constraints on review capacity.

Broad Stakeholder Discussions Could Help Coordinate Use of Natural Language Processing for Patient Narratives

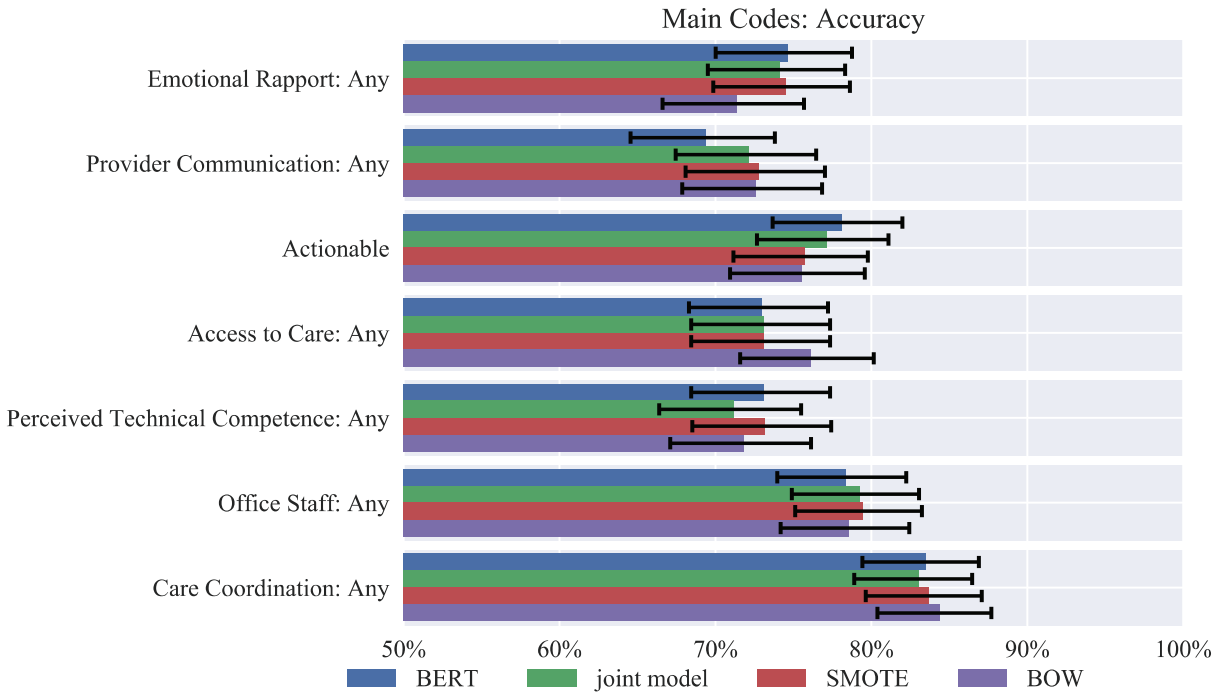
Health care providers and organizations tasked with improving the quality of health care could also consider organizing ongoing broader discussions among stakeholders about the role of natural language processing in analyzing health care data. The data used here come from standardized surveys administered to systematically sampled groups of health care patients, which has many advantages for developing common analytic approaches. Logical extensions should consider other, less structured narrative data sources, such as online rating sites and statements on Twitter. Once priorities are identified, a public benchmark that measures the ability of a model to achieve those priorities and associated dataset could be produced and released. Public benchmarks, such as GLUE and SuperGLUE for natural language understanding and ImageNet for image recognition, have a history of focusing and driving innovation in the ML community.[19] In addition to possibly driving interest from the academic ML community and even leading to high-performance open-source models, such a benchmark would provide a common language for evaluating the performance of proprietary models.

[19] For more information on these benchmarks, see General Language Understanding Evaluation Benchmark, undated; SuperGLUE Benchmark, undated; and ImageNet, undated.

Appendix: Additional Results

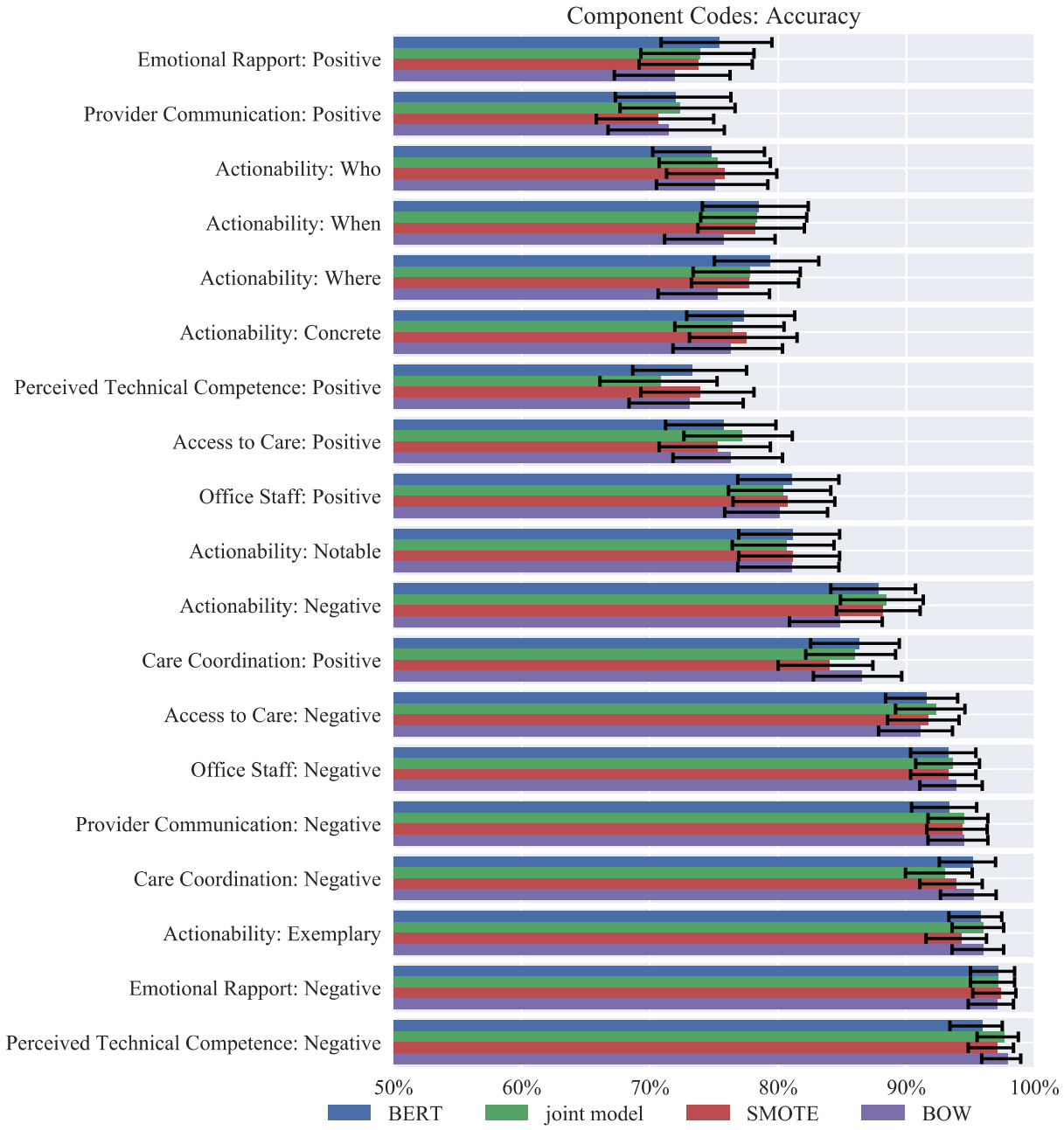
This appendix contains additional data that did not merit inclusion in Chapter 4, as it did not materially alter the conclusions drawn in that chapter. Figures A.1 and A.2 depict the accuracies of all four models on the test set. As mentioned in the Chapter 4, the four models performed similarly as measured by accuracy. In addition, for most codes the performance of the four models appeared qualitatively similar, and no systematic differences stand out among the four models considered.

Figure A.1. Accuracy Performance on Presence Versus Absence of Any Mention of Each Code in All Four Models



NOTE: Error bars represent 95-percent confidence intervals.

Figure A.2. Accuracy Performance on Presence Versus Absence of Subcodes in All Four Models.

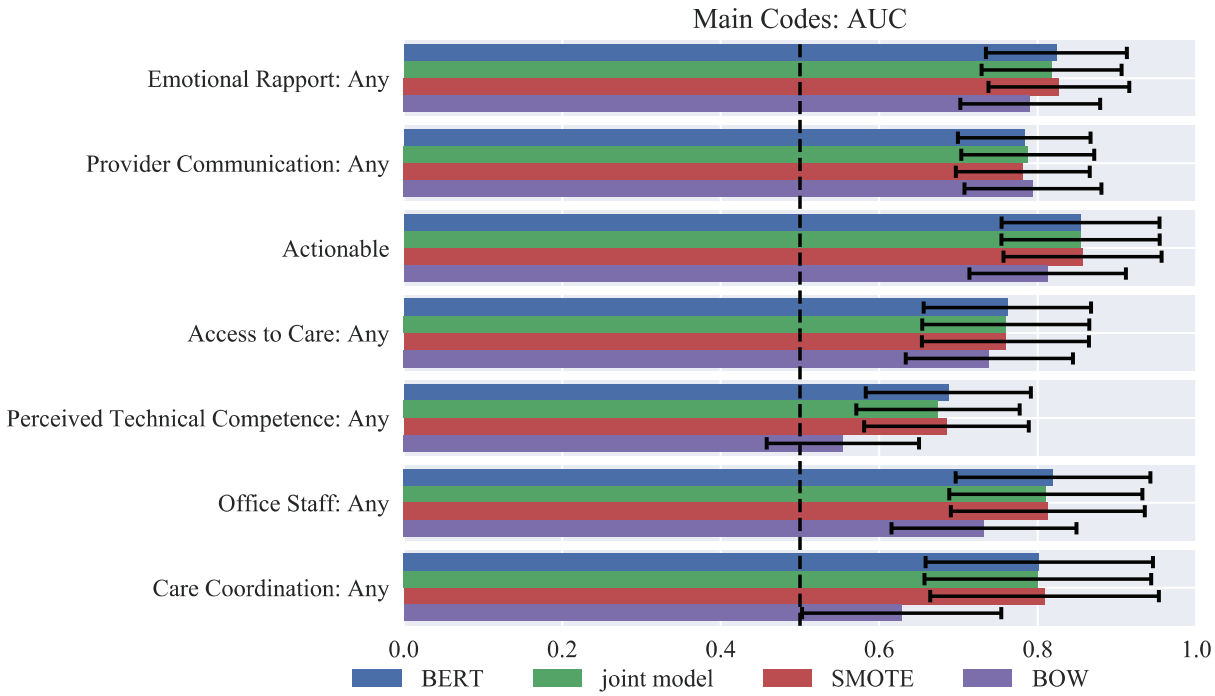


NOTE: Error bars represent 95-percent confidence intervals.

Here, we also report the AUC for our final fitted models. The AUC inspects the internal structure of the model, utilizing a *score* produced by the model for each narrative that forms the basis of the model's classification decision. The score represents the model's degree of certainty that the narrative contains the code. Unlike the final classification decision, the score does not explicitly consider the assumed operating condition. The AUC uses the test set to compute the

probability that a random positive example will have a higher score than a random negative example. We include the AUC as a measure of classifier power that is independent of operating condition. Caution is indicated when interpreting the AUC, however, as the higher AUC for model A than model B does not imply that model A will outperform model B in all operating conditions. Here, the baseline is 0.5, which represents the performance of random guessing (DeLong et al., 1988).

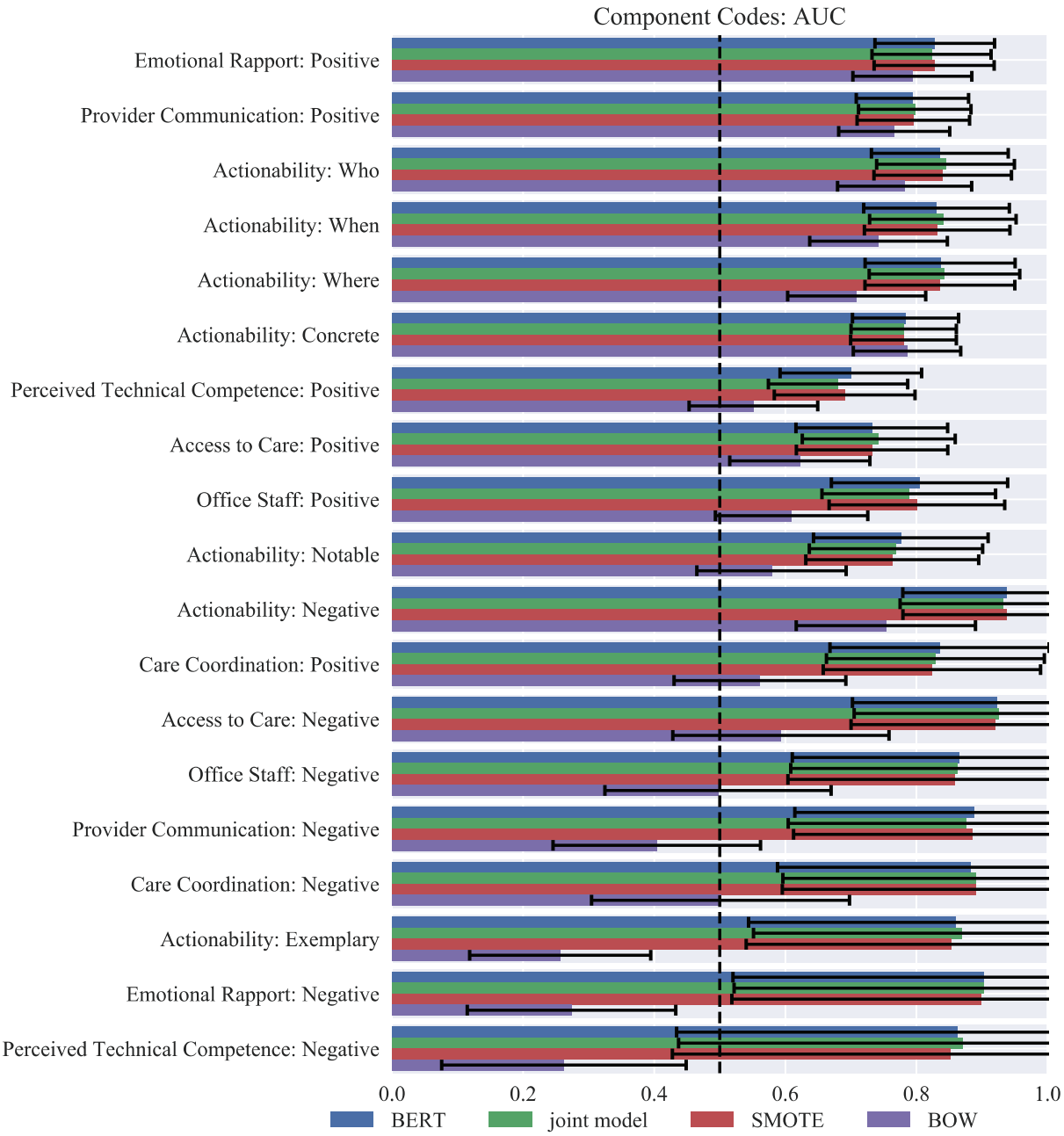
Figure A.3. Area-Under-the-Curve Performance on Presence Versus Absence of Any Mention of Each Code in All Four Models.



NOTE: The vertical line at 0.5 represents performance from random guessing. Error bars represent 95-percent confidence intervals.

Generally, the AUC supports the conclusions about the relative efficacy of the simple embedding and the BERT embedding reached by considering the balanced accuracy.

Figure A.4. Area-Under-the-Curve Performance on Presence Versus Absence of Subcodes in All Four Models



NOTE: The vertical line at 0.5 represents performance from random guessing. Error bars represent 95-percent confidence intervals.

References

Alkahtani, H. S., Gardner-Stephen, P., & Goodwin, R. (2011). A taxonomy of email SPAM filters. In *The 12th International Arab Conference on Information Technology*, 351–356, Riyadh, Saudi Arabia.

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, *38* (5), 2916–2957.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16* (3), 199–231

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16* (2), 101–117.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA*, *318* (6), 517–518.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9* (2), 48–57.

Casella, G., & Berger, R. L. (2002). *Statistical inference*, Vol. 2. Pacific Grove, Calif.: Duxbury.

Castelvecchi, D. (2010). Can we open the black box of AI? *Nature*, *538* (7623), 20–23.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., . . . & Sung, Y. H. (2018). Universal sentence encoder. As of September 10, 2020: arXiv preprint arXiv:1803.11175

Chawla, N, V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, *37* (1), 51–89.

Creswell, J. W., & Poth, J. (2018). *Qualitative inquiry & research design—Choosing among five approaches*, 4th ed. London: Sage.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44* (3), 837–845.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132* (20), 1920–1930.

De Smet, F., Moreau, Y., Engelen, K., Timmerman, D., Vergote, I., & De Moor, B. (2004). Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer, 91*, 1160–1165.

Devlin, J., Chang, M-W, Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. As of September 10, 2020: arXiv:1810.04805v2

Dias, R. D., Gupta, A., & Yule, S. J. (2019). Using machine learning to assess physician competence: A systematic review. *Academic Medicine*, *94* (3), 427–439.

Divita, G., Carter, M., Redd, A., Zeng, Q., Gupta, K., Trautner, B., . . . & Gundlapalli, A. (2015). Scaling-up NLP pipelines to process large corpora of clinical notes. *Methods of Information in Medicine*, *54* (6), 548–552.

Elkan, C. (2001) The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Vol. 2, 973–978. San Francisco: Morgan Kaufmann Publishers, Inc.

Elkin, P. L., Froehling, D., Wahner-Roedler, D., Trusko, B., Welsh, G., Ma, H., . . . & Brown, S. H. (2008). NLP-based identification of pneumonia cases from free-text radiological reports. In *American Medical Informatices Association Annual Symposium Proceedings*, Vol. 2008, 172–176. Bethesda, Md.: American Medical Informatics Association.

FitzHenry, F., Murff, H. J., Matheny, M. E., Gentry, N., Fielstein, E. M., Brown, S. H., . . . & Speroff, T. (2013). Exploring the frontier of electronic health record surveillance: The case of postoperative complications. *Medical Care*, *51* (6), 509–516.

Flach, P. A. (2003). The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proceedings of the 20th International Conference on Machine Learning*, 194–201, Washington, D.C.

Fowler, F. J., Jr., Cosenza, C., Cripps, L. A., Edgman-Levitan, S., & Cleary, P. D. (2019). The effect of administration mode on CAHPS survey response rates and results: A comparison of mail and web-based approaches. *Health Services Research*, *54* (3), 714–721.

Friedman, C., Rindflesch, T. C., & Corn, M. (2013). Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of Biomedical Informatics*, *46* (5), 765–773.

General Language Understanding Evaluation Benchmark (undated). As of September 10, 2020: https://gluebenchmark.com

GLUE—*See* General Language Understanding Evaluation.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, *15* (11), e239.

Grob, R., Schlesinger, M., Barre, L. R., Bardach, N., Lagu, T., Shaller, D., . . . & Palimaru, A. (2019). What words convey: The potential for patient narratives to inform quality improvement. *Millbank Quarterly*, *97* (1), 176–227.

Grob, R., Schlesinger, M., Parker, A. M., Shaller, D., Barre, L. R., Martino, S. C., . . . & Cerully, J. L. (2016). Breaking narrative ground: Innovative methods for rigorously eliciting and assessing patient narratives. *Health Services Research*, *51* (Suppl 2), 1248–1271.

Gu, Y., & Leroy, G. (2020). Use of conventional machine learning to optimize deep learning hyper-parameters for NLP labeling tasks. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 1026–1035, Maui, Hawaii.

Guetterman, T. C., Chang, T., DeJonckheer, M., Basu, T., Scruggs, E., & Vydiswaran, V. G. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research*, *20* (6), e231.

Hays, R. D., Berman, L. J., Kanter, M. H., Hugh, M., Oglesby, R. R., Kim, C. Y., . . . Brown, J. (2014). Evaluating the psychometric properties of the CAHPS patient-centered medical home survey. *Clinical Therapeutics*, *36* (5), 689–696.

Huppertz, J. W., & Otto, P. (2018). Predicting HCAHPS scores from hospitals' social media pages: A sentiment analysis. *Health Care Management Review, 43* (4), 359–367.

Huppertz, J. W., & Smith, R. (2014). The value of patients' handwritten comments on HCAHPS surveys. *Journal of Healthcare Management*, *59* (1), 31–48.

ImageNet (undated). As of September 10, 2020: http://www.image-net.org

Iyer, S. V., Harpaz, R., LePendu, P., Bauer-Mehren, A., & Shah, N. H. (2014). Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association*, *21* (2), 353–362.

Jha, A. K. (2011). The promise of electronic records: Around the corner or down the road? *JAMA*, *306* (8), 880–881.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, *14* (2), 1137–1145.

Kuckartz, U. (2014). *Qualitative text analysis: A guide to methods, practice and using software*. Thousand Oaks, Calif.: SAGE.

Leroy, G., Gu, Y., Pettygrove, S., Galindo, M. K., Arora, A., & Kurzius-Spencer, M. (2018). Automated extraction of diagnostic criteria from electronic health records for autism spectrum disorders: Development, evaluation, and application. *Journal of Medical Internet Research*, *20* (11), e10497.

Lopez, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What patients say about their doctors online: A qualitative content analysis. *Journal of General Internal Medicine*, *27* (6), 685–692.

Martino, S. C., Shaller, D., Schlesinger, M., Parker, A. M., Rybowski, L., Grob, R., . . . & Finucane, M. L. (2017). CAHPS and comments: How closed-ended survey questions and narrative accounts interact in the assessment of patient experience. *Journal of Patient Experience*, *4* (1), 37–45.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119, Neural Information Processing Systems Foundation, Inc., San Diego, Calif.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*, 3rd ed. London: Sage.

Murff, H. J., FitzHenry, F., Matheny, M. E., Gentry, N., Kotter, K. L., Crimin, K., . . . & Speroff, T. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, *306* (8), 848–855.

National Academies of Sciences, Engineering, and Medicine (2019). *Implications of artificial intelligence for cybersecurity: Proceedings of a workshop*. Washington, D.C.: The National Academies Press. As of September 10, 2020: https://www.nap.edu/catalog/25488/implications-of-artificial-intelligence-for-cybersecurity-proceedings-of-a-workshop

Nawab, K., Ramsey, G., & Schreiber, R. (2020). Natural language processing to extract meaningful information from patient experience feedback. *Applied Clinical Informatics*, *11* (2), 242–252.

Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., . . . & Merchant, R. M. (2016). Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs*, *35* (4), 697–705.

Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W., & Johnson, K. B. (2011). Data from clinical notes: A perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, *18* (2), 181–186.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, *9* (2), 65–78.

Salt, J., Harik, P., & Barone, M. A. (2019). Leveraging natural language processing: Toward computer-assisted scoring of patient notes in the USMLE Step 2 Clinical Skills Exam. *Academic Medicine*, *94* (3), 314–316.

Schlesinger, M., Grob, R., Shaller, D., Martino, S. C., Parker, A. M., Rybowski, L., Finucane, M. L., & Cerully, J. L. (2020). A rigorous approach to large-scale elicitation and analysis of patient narratives. *Medical Care Research & Review, 77* (5), 416–427.

Scholle, S. H., Vuong, O., Ding, L., Fry, S., Gallagher, P., Brown, J. A., Hays, R. D., & Cleary, P.D. (2012). Development of and field-test results for the CAHPS PCMH survey. *Medical Care*, *50* (Suppl), S2–10.

SuperGLUE Benchmark (undated). As of September 10, 2020: https://super.gluebenchmark.com

Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., . . . & Si, L. (2019). Structbert: Incorporating language structures into pre-training for deep language understanding. As of September 10, 2020: arXiv preprint arXiv:1908.04577

Wang, X., Hripcsack, G., Markatou, M., & Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *Journal of the American Medical Informatics Association*, *16* (3), 328–337.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. As of September 10, 2020: arXiv:1509.01626v3