2021

Promises and Pitfalls of Machine Learning Classifiers for Inter-Rater Reliability Annotation

Lucille Dorothy Ayres
Wright State University

PROMISES AND PITFALLS OF MACHINE LEARNING CLASSIFIERS FOR INTER-RATER RELIABILITY ANNOTATION

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

DOROTHY LUCILLE AYRES

B.S., Wright State University, 2018

2021

Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

APRIL 26, 2021

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Dorothy Lucille Ayres ENTITLED Promises and Pitfalls of Machine Learning Classifiers for Inter-Rater Reliability Annotation BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Valerie Shalin, Ph.D.
Thesis Director

David Lahuis, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Committee on
Final Examination

Debra Steele-Johnson, Ph.D.

T. K. Prasad, Ph.D.

Barry Milligan, Ph.D.
Vice Provost for Academic Affairs
Dean of the Graduate School

ABSTRACT

Ayres, Dorothy Lucille. M.S., Department of Psychology, Wright State University, 2021.
Promises and Pitfalls of Machine Learning Classifiers for Inter-Rater Reliability Annotation

Qualitative data result from observation, video, and dialogue. These types of data are flexible and allow us to study behavior without imposing potentially disruptive data collection methods. However, subsequent quantitative analysis requires a time consuming, labor intensive initial coding process, and a second manual coding to calculate inter-rater reliability. I examined the use of machine learning algorithms to reduce the amount of manual annotation work required to perform inter-rater reliability measures on text data. By comparing machine-human and human-human raters using Cohen's Kappa statistic and an informal analysis of the features used in machine learning classification, I identify the promise and limitations of machine rating for conducting the second coding effort used to determine reliability. I found that machine learning algorithms can be useful tools for supporting inter-rater reliability as a second coder, but there are limitations associated with the class balance of the data that may restrict their usage.

Table of Contents

List of Tables

Table

Acknowledgements

First and foremost, I need to thank my thesis supervisor, Valerie Shalin, and members of the WSU Computer Science department Shreyansh Bhatt, Swati Padhee, and Manas Gaur. Without their continual assistance and advice during my work on this thesis, I would not have been able to complete this project. I would like to thank you all very much for your understanding and help over the years.

I would also like to thank my committee, including Valerie Shalin, Debra Steele-Johnson, and T. K. Prasad. As my advisor, Dr. Shalin supported me in this project every step of the way and this support was crucial to my work. Dr. Steele-Johnson expressed an interest in my project early on, and has continually brought a fresh perspective from her Industrial/Organizational psychology background that helped me clarify many ideas in this thesis. Dr. Prasad helped me understand the computer science perspective and was always willing to explain how and why CS projects use machine learning tools, which was vital to this work. Thank you all so much.

Additionally, I would like to thank the other members of the Workplace Cognition lab at Wright State, as well as Scott Watamaniuk. The other members of the lab provided important social support as well as being available for conversations and questions about both my project and more general academic questions. Dr. Watamaniuk, as my undergraduate thesis advisor, played a vital role in my academic career. Thank you all for your support and time.

Finally, I would like to extend my thanks to my friends and family, who have supported me more than I can express. Thank you to my parents and grandparents, without whose support I could not have made it this far. I have experienced some ups and downs in the past few years, and without the support of the people closest to me I would never have gotten this far. Thank you, to everyone.

Introduction

Real-world observational studies rarely permit the systematic collection of content-oriented quantitative behavioral metrics. Such studies rely on *post hoc* human annotation of qualitative data that result from observations, video and dialogue. However, behavioral science favors quantitative measures over the social scientist's qualitative summaries. This is one motivation for the controlled laboratory study where the collection of quantifiable metrics is built *a priori* into the data collection method. Relative to quantitative measures, qualitative data do not require specific formats or participant inputs. This makes these measures particularly useful for psychological research outside of the laboratory. However, the flexibility comes with a coding cost to support quantitative analysis. Manual coding (also called annotation) is time-consuming and expensive. Once manually coded initially, a second re-coding examines the consistency of the coding scheme. This step is important to determine whether the coding is reliable. If it is not, researchers can not determine whether their data supports their claims or whether the categories they have defined are repeatable.

A potential automated option for the second manual coder in inter-rater reliability is machine learning (ML). These programs develop classification rules that sort initially coded instances into categories. This study attempts to substitute ML technology for a second human rater to make the inter-rater reliability coding process faster, more efficient and itself more reliable. In the remainder of this introduction I define machine learning and the rationale for using it in this project. I then discuss potential issues with using ML for inter-rater reliability, as well as the methods used for calculating inter-rater reliability in the first place. I complete this introduction with a discussion on using ML as a second rater in inter-rater reliability calculations.

The Methodological Problem

Auerbach and Silverstein (2003) defined qualitative research as “research that involves analyzing and interpreting texts and interviews in order to discover meaningful patterns descriptive of a particular phenomenon”. Some examples of qualitative data include interviews, histories, text analysis, and descriptive ethnographies. Qualitative data often comes in the form of unstructured text. Psychologists typically code, or annotate such data to convert it to metrics suitable for quantitative analysis.

Current Annotation Approaches

Social and behavioral science employs different qualitative research methods with corresponding approaches to annotation. In his book *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* (2007) the social scientist John W. Creswell distinguishes several varieties of qualitative research on text-based data. In *narrative research*, researchers gather stories from an individual or a few individuals and then “re-story” them, imposing chronological order, adding causal links, and exposing dichotomies, sometimes in collaboration with the source as part of identifying key topics. In *phenomenological research*, researchers interview several subjects regarding their experience of a common phenomenon that may transcend cultures (e.g., grief, medical intervention) and interpret the transcripts of those interviews for commonalities in “what” the participants experienced and “how”, to obtain a single unified description of the phenomenon. *Ethnographic research* is more ambitious than phenomenological research by encompassing an entire cultural group over time. Researchers conducting ethnographic studies gather data from a variety of sources, including interviews and observational fieldwork. Researchers then attempt to organize that data into a coherent understanding of the culture’s norms, values and constraints, sometimes employing a critical stance on the role of power and the marginalization of the powerless. *Case study research* is more bounded than an ethnography, and examines one or a few specific cases of a phenomenon in-depth by gathering information from multiple sources. Those sources can include interviews, text recorded observations, video and audio recordings, and documents. The researchers then put the data in chronological order and identify themes that indicate the complexity of the case, including its context and causal relationships.

Less broad than ethnography, *grounded theory research* extends phenomenology with a new unified theory about a phenomenon grounded in the new data, often collected in the field by observing behavior in its social setting. Using this approach, the primary issues are "saturation", i.e., a sufficient number of observations (20 -30), expanded sampling after initial analysis (up to 60), and of particular importance here, the development of the coding scheme itself. First researchers use *open coding,* where they create categories into which they sort the data, such as causes, strategies, context and outcomes. After open coding, a researcher employs *axial coding* to organize those categories into some type of logical relationship.

Grounded theory best captures the approach used to code the data used in this study (Robinson et al., 2020). Robinson suggested a number of independent categories associated with physician strategies for processing patients in the emergency room. For example, the information gathering category requires features such as "current symptoms", referring to the type of information that the doctor is gathering, and "the patient", referring to the source of the information. He intentionally avoided complaint-specific causes to preserve generalizability, and did not have access to outcomes (as is typical in this setting). When creating a coding scheme, psychologists like Robinson seek themes that are key to the psychological constructs, e.g., workload or conscientiousness. The resulting coding scheme can be quite abstract, requiring substantial inference and interpretation between the actual data (e.g., lexical items in the text) and the category in question. An example of the qualitative data from the present study is "res asks about any belly surgery" coded as an information gathering behavior. "Asks" is likely a relevant feature but it is not the only feature indicative of information gathering.

Problems with the Grounded Theory Annotation Scheme

Typically the psychologist creates a new, theoretically informative coding scheme and class definitions for abstract constructs. The same researcher also conducts the first classification of the data. However, the researcher's coding scheme and definitions may not correspond to reproducible categories. The experimenter could be a biased rater, finding patterns that cannot be reproduced. To address this risk, psychologists conduct inter-rater reliability. Technically, inter-rater reliability is the quantification of agreement between different raters when rating the same data using a detailed, pre-established coding scheme. Replicability confers precision in the coding scheme. For example, two raters with good reliability would rate one piece of text as representing the same type of behavior. Typically, researchers specify the rules and associated features that define which category an instance represents; for example "asking a question" may define the "information gathering" category. Then either new raters attempt to apply these definitions to a subset of the data or the original rater applies these rules after significant time delay when memory of the initial instance and its coding have decayed. High inter-rater reliability occurs when there is a high level of calculated agreement between raters on the classification of individual items in the data set. Because the definitions of the categories are established in advance, the assumption is that independent review is applying the same feature set (Kottner et al., 2011; Burns, 2014). While replicability confers precision, it does not confer accuracy, that is proximity to truth.

Problems with the Annotation Process

The large sets of qualitative data required for statistical analysis increase the demand on annotation. The labor intensive process of both annotating that data and assuring coding scheme reliability not only delays subsequent analysis, conclusions, and publication (Davidson et al., 2018) but in so doing limits the kinds of studies that can be pursued. Additionally, human raters are fallible. Human raters can suffer from drift over time, particularly with large data sets. Accordingly, their criteria for rating potentially changes over time, and therefore the reliability of their ratings across the whole data set degrades. Moreover, annotation is similar to a signal detection task, with known vulnerability to class imbalance (Pandey et al., in review). When one class is under-represented, a human annotator may have difficulty identifying it, or may make implicit adjustments in the decision rule. A small number of positive instances causes viewers to raise their criterion, while a large number of positive instances causes viewers to lower their criterion.

Because of concerns for both the annotation scheme and the annotation process as noted above, psychological research does not treat the initial human rating as a gold standard, particularly when it is based on categories constructed by a single observer. For these reasons, psychologists seek inter-rater reliability by enlisting a second observer. Of course, the second observer is subject to the same limitations as the first, and will be unable to repeat an imprecise coding scheme.

Machine Learning Solutions

In this thesis, I compare the agreement of a machine learning classifier with the original human annotator and the agreement of two human annotators to determine whether ML provides a reasonable substitute for a second human coder in typical psychological inter-rater reliability.

Computer science researchers use so-called supervised ML methods to classify manually annotated qualitative data. Applications include videos of lecturers (Brooks, Amundson & Gree, 2009), recordings of human speech (Kang & Johnson, 2015) and medical data (Williams, Weakley, Cook & Schmitter-Edgecombe, 2013). In the medical field, ML identifies the features to diagnose dementia (Shankle, Mani, Dick & Pazzani, 1998). Medical researchers also use it to predict dementia scores based on patient data (Williams, Weakley, Cook & Schmitter-Edgecombe, 2013). In non-medical fields, researchers use ML algorithms to search for important moments in video recordings of lectures, with the possibility of improving online education outcomes (Brooks, Amundson & Greer, 2009).

In the remainder of this section, I will discuss the definition of machine learning and my rationale for using it, potential issues with using machine learning, calculating inter-rater reliability, and using machine learning to replace the second rater.

Definition and Rationale for Machine Learning

In supervised learning, the ML algorithm employs a subset of human annotated cases to develop its own classification rules for the features of instances that determine human annotation. In computer science, human ratings are typically treated as the gold standard. The ultimate goal is to have the machine replicate the human approach. In one popular method, ML algorithms create something called a decision tree; each time the algorithm identifies a feature in a piece of text, it must decide whether that feature is an indicator of the category or not. As it examines more features, the algorithm creates a set of decision rules regarding whether a certain feature is indicative of the category. For example, the program might identify "asks" as a diagnostic feature of "information gathering". A good machine learning model might indicate correlated features missed in the explicit category definitions.

In the conventional ML application, a bad computer model results from concepts too difficult for the machine to acquire, most typically due to features accessible to the annotator that are not accessible to the machine. Voting resolves differences between multiple annotators. Low human agreement excuses machine failure and the machine is not expected to exceed human agreement or repair human rater variability. On the other hand, psychologists would also suggest an initially bad classification scheme.

However, good agreement between the machine and the annotators suggests that the annotators are responding to a well-specified coding scheme that the machines can apprehend and mimic.   A good model has the potential to replace manual annotation with machine annotation. This could reduce the manual labor required to perform inter-rater reliability, enable the examination of large data sets and facilitate emergent coding that tracks thematic changes in unstructured data over time. Moreover, once an acceptable automated coding model has been validated by testing the ML algorithm on a reserved set of data, automated coding could be used for the purpose of annotating future collected data. Nevertheless, the absence of a reliable coding scheme is just one of the explanations for bad ML results. Hence bad results do not necessarily condemn the coding scheme.

This experiment reverses the conventional Computer Science approach to annotation. I consider that the machine algorithm may be more objective, be able to identify specific features in the data that human raters cannot, and would not suffer from rating drift over time. Given a quality training data set, an ML algorithm could potentially surface hidden in the data that are not apparent to human raters and perform more consistently than a human rater. On the other hand, and as discussed further below, humans are better aware of the intent and context in the data, and therefore better able to discriminate meaningful "causative" patterns from spuriously correlated content.

Potential Issues With Machine Learning as a Second Rater

The successful application of ML as a second rater is vulnerable to a number of limitations. If behavioral researchers want to use this tool, we must examine whether ML programs can match human annotators in both reliability and content. In addition, selecting a unit of analysis in unstructured text data falls outside the guidelines of ML methods, and yet must be established.

Ideally the features that the algorithm finds to categorize the data correspond to the pre-established annotation rules. However, the program may not capture the same features as human raters who might respond differently to feature combinations or context (Kang & Johnson, 2015). In the case of a mismatch, the algorithm might be overfitting, in which idiosyncratic features of the data set (e.g. *belly*, or *surgery* above) contribute to reliability statistics (e.g., a high Kappa value) but will not generalize to a different corpus, thereby limiting utility. This concern echoes the anthropologists' saturation caution noted above, regarding the breadth of the sample submitted to analysis. A related concern is the dependence of ML algorithms on word frequency, neglecting low frequency, but discriminating words in favor of high frequency, potentially spuriously correlated high frequency words. For this reason, computer scientists often eliminate so called stop-words, that is, high frequency words, such as articles and pronouns.

Machine learning algorithms frequently struggle in cases of class imbalance. When one class dominates the data set, the algorithm may simply label all cases as that class and still be correct a majority of the time. SMOTE (Synthetic Minority Oversampling Technique) is one method to address the problem of overfitting due to class imbalance. SMOTE is an R function that oversamples from a minority class. Oversampling from the minority class artificially creates a better class balance so the class imbalance does not throw off the ML algorithm (Maciejewski & Stefanowski, 2011). A more sophisticated alternative to SMOTE is GANS (Generative Adversarial Networks) which combines a generative model and a discriminatory model to replicate patterns artificially in the data with new instances (that come from the same distribution but are not duplicates).

Another consideration is the unit of analysis of episodes for classification. Machine Learning uses a binary classification of an episode---it either is or is not an instance of the class no matter how many features or separate indicators are present. However, continuous observations lack predefined units of analysis. Temporally based units (e.g., delineated by minutes) may still contain multiple pieces of evidence concerning the classification of that unit. Content based units (e.g., turns in a conversation) that divide the original data require manual determination. I return to this issue below regarding the proposed data set.

Inter-Rater Reliability Calculation

Psychologists and Computer Scientists differ in the way they quantify agreement. The Computer Scientist typically quantifies agreement between the human annotator and the ML classifier based on positive and negative agreement. Two scores contribute to agreement. Recall concerns the proportion of positive instances captured by the classification scheme. A low recall value means that many positive instances were missed. Precision concerns the proportion of instances classified as positive were manually classified as positive. A low precision value implies a high false alarm rate. The F measure combines these two metrics to better mirror the tradeoffs between them.

Psychologists usually quantify inter-rater reliability in qualitative data categorization with Kappa statistics, although alternatives such as Gwet's AC1 exist. The multiple versions of Kappa statistics include Cohen's Kappa, Cohen's weighted Kappa, and the intraclass Kappa statistic (Kottner et al., 2011). I am using Cohen's Kappa for this study. Low values of Cohen's Kappa indicate poor agreement between annotators. Sim and Wright (2005) described the purpose of Kappa is to measure "true" agreement between two researchers' use of the same analytic tool. By "true" they mean that Kappa discounts agreement that occurs merely through chance. Researchers use Kappa to measure the proportion of agreement between two raters beyond chance. The equation for Kappa is

$$K = \frac{(P_o - P_c)}{(1 - P_c)}$$

where P_o is the proportion of observed agreements and P_c is the expected proportion of chance agreements. Kappa can range from -1 to 1, and a result of zero indicates that there was no more agreement than could be expected by chance. The closer Kappa is to 1 or -1 the more unity there is between the raters. By accounting for chance, Kappa takes into account class imbalance in the data. That is, if the data contains instances of one category more than another, an ideal "guess" merely reflects the extent of the imbalance. For example if the data contain 80% instances of "A" and 20% instances of "~A", the best guess is "A" and agreement will appear to be relatively high without any motivating classification analysis. Kappa protects against this kind of spurious agreement by assessing it over and above the agreement we expect by chance. Class imbalance increases with departures from an equal distribution of instances across the categories.

Using standard criteria, a Kappa value above 0.75 indicates a strong reliability (agreement) between the machine and human categorization. A Kappa value between 0.40 and 0.75 indicates a moderate reliability, while a Kappa below 0.40 indicates a weak reliability

(Fleiss, 1981). Landis and Koch (1977) suggest more levels of agreement: 0.0 to 0.20 indicates slight agreement, 0.2 to 0.4 indicates fair agreement, 0.41 to 0.60 indicates moderate agreement, 0.61 to 0.80 indicates substantial agreement, and >0.80 indicates almost perfect agreement. Conventionally, publication of research in psychological journals hinges on obtaining at least moderate agreement.

Using Machine Learning as a Second Rater

The data I am using for this study was classified against a number of independent categories, all of which had defined features. For example, the information gathering category requires features such as "current symptoms", referring to the type of information that the doctor is gathering, and "the patient", referring to the source of the information. The researcher uses these features to identify whether a piece of text belongs in a category or not.

The goal in this research was not to optimize the ML algorithm, but to examine the potential utility of an existing ML algorithm as a second rater. I sought to compare the agreement between the human annotators, between the human and the machine annotators and whether the human and machine annotations use similar terms to classify the data. For this initial investigation, I tested the Random Forest algorithm, as implemented in the Waikato Environment for Knowledge Analysis, or WEKA. I listed the steps that I used to prepare the data and analyze the data in WEKA in Appendix B. According to Hall et. al., (2009), researchers at the University of Waikato in New Zealand initiated WEKA in 1992 to provide a toolbox of learning algorithms for researchers as well as an environment they could use to create their own algorithms. WEKA is now a widely-used toolkit that includes several machine learning algorithms and other data preprocessing tools, including algorithms for categorization and attribute selection.

The Random Forest algorithm is not the only possibility. The K-nearest neighbors algorithm operates on a vector representation of the text data, classifying an instance according to its similarity to its neighbors. The K-nearest neighbors algorithm is vulnerable to over-fitting, and its vector representation does not support meaningful inspection. A Linear Classifier seeks a linear combination of features that identifies class membership. A hyperplane separates positive and negative instances. A Support Vector Machine is a particular type of linear classifier that finds the hyperplane that maximizes the distance between the positive and negative instances. It can also be used to deal with non-linear class boundaries using data transformation, but this is beyond the scope of this work.) The Random Forest algorithm combines a set of feature-based decision trees to classify an instance according to the central tendency of the trees. It is a simple approach, and it has the advantage of generating an inspectable forest of decision trees as part of the final output.

Research Questions

Agreement between the human and the machine annotation suggested that ML algorithms can potentially substitute for a human annotator. It was also important to check the features the program uses because that will determine how broadly I can apply an algorithm after training. I may have needed to alter the data, including potentially removing stop words such as "and" or "but" to prevent the algorithm from catching them as significant when they are not useful.

To address the promise of machine learning for inter-rater reliability I posed the following questions and corresponding analyses for the data set in question:

RQ1: What is the quantified inter-rater reliability between the human and machine annotators?

RQ2: How do the machine learning categorization terms compare with the original class definitions?

RQ3: Do I need to eliminate stop words (highly frequent, typically non-discriminators e.g., "the") from the data set to suppress spurious feature selection for qualitative data classification?

RQ4: What is the effect of drift in the human annotation process?

RQ5: What is the quantified inter-rater reliability between two human annotators?

RQ6: How does class imbalance affect machine learning and human annotation?

Pilot Results

Pilot work examined the information gathering category, which was the most balanced in the data set. When comparing the human and machine annotators on “information gathering”, I found a Kappa of 0.76 showing a strong agreement. The resulting ML model also chose terms that are associated with the original classification rules of this data. For the information gathering condition, the most important terms it selected were terms like “asks”, “does”, “says”, “looks”, “feels”, and “listens”. The pilot WEKA results including the decision tree terms appear in Appendix C. These terms were also associated with things that a human annotator might seek to see if a described behavior is for the purpose of gathering information. This suggests that a machine learning algorithm can be relied on to give results in line with human annotation and can be used as a substitute for a second human rater.

Method

Data Set

Robinson (2011) provided the data for this project (see Appendix A). This data consists of descriptions of doctor behavior observed over the course of a work shift. It was taken from twenty-six emergency department physicians at two different hospitals. It included a total of 38 excel files. Each file was composed of observations of one physician over the course of one shift including several hundred instances of data with one instance per row. In all, these files included approximately 25,000 separate entries. I do not know the specific criteria the original observer used to separate the data into individual instances, but it may have been something like pauses in the action by the physicians or breaks between statements. Some examples of instances include “asks what happened to bring them in”, “says they've already ordered a chest pain set and BNP”, and “looks at pt's mouth and eyes”.

Coding Scheme

The original annotator, Eric Robinson, categorized these descriptions into six different broad behavior types (see Table 1). These categories were: information gathering behavior, diagnostic behavior, evidence evaluation behavior, patient management behavior, system management behavior and filtering behavior. For my final analysis I chose three specific categories of behavior and ran them through WEKA separately in order to compare the results between them. The three categories I chose were information gathering from a patient source (PT), system management (SM), and logistics (LOG), which was a subcategory of patient management. These categories provided examples of data that is well balanced (26% positive instances), moderately balanced (15% positive instances) and poorly balanced (6% positive instances), respectively. I also chose them because the researcher who collected this data found that these three variables (unlike several others) were relevant to his original analysis. Relative to the pilot study, information gathering from a patient source (examined in the present analysis) is a subset of the complete information gathering category. The text data itself is exactly the same for all variables; the only thing that changes is the specific type of behavior being categorized.

According to the Robinson coding scheme, doctors use information gathering behavior to generate facts about the case. This information is further sub-categorized into types of information and sources of information such as information gathered from the patient or from medical records. Diagnostic behaviors are any behaviors that attempt to find a cause for symptoms or to eliminate possible causes. Evidence evaluation behaviors help doctors determine which signs and symptoms are important and worth pursuing. Patient management behaviors include any actions the doctors take to treat their patients and also to keep them informed and to act in their best interest. System management behaviors include any behaviors doctors use to maneuver through the larger system of the hospital and health care system they operate within. Filtering behaviors are actions doctors take to limit the number of problems they must handle. Any data segment could pertain to multiple data categories. For this reason, I ran separate classification models with the same data set.

Table 1. Robinson Analysis Categories

Behavioral Categories	**Potential Indicators of Behavior**
Information Gathering Subcategories:	Questions, looking up records, searching databases

● Source (Exams, tests/images, **the patient (PT)**, family/friends, medical records, hospital staff, internet/references, misc.) ● Type (Current symptoms, timeline, past medical info, contributors, reference, other)	
Diagnostic	Using treatments to diagnose the issue
Evidence Evaluation	Checking the reliability of the patient history
Patient Management ● Type (Collaboration, treatment, consulting, **logistics (LOG**))	Keeping the patient informed and reassured, regulating the patient's condition, offering advice on care
System Management (SM)	Contacting a specialist, working in a busy environment
Filtering	Making judgements on which problems to address

Note. Categories used in this thesis project have been bolded.

Instances

Researchers often have access to text data in a continuous stream format. Continuously streamed data raise a problem with the unit of analysis. The text does not necessarily have principled delineation of discrete instances for subsequent annotation. The data that I am using is broken into smaller parts, but those parts are not standardized and do not each exhibit only one type of behavior. For example, the piece of text data “res tells pt the xray and US were ok, says they're waiting on the blood work; asks if the pain meds helped” was originally annotated as containing multiple instances of evidence evaluation, information gathering, and system management behavior.

One potential solution could be to cut the existing data into single units so that every unit only contains one instance of a behavior type. However, this solution would require an initial manual segmentation of the data, inconsistent with the goal of streamlining manual annotation. Another solution is to remove observations that are classified as containing multiple instances of behavior. This could work but depends on the number of observations that contain multiple instances and the class imbalance concern. A viable solution is re-classifying each entry into the data file as either zero or one. That is, any classification above one becomes one, indicating that piece of text does include at least one instance of that type of behavior, and zeros remain zeros, indicating that text does not contain that type of behavior. This solution is consistent with standard practice in the automated analysis of unstructured short social media posts, and can be applied to a variety of different types of data (Bhatt et al., 2019; Kursuncu et al., 2019; Purohit et al., 2013; Purohit et al., 2014). Therefore it is the solution I used for this experiment. Doing so resulted in 25,480 instances.

Recoding Procedure

I re-coded 10 percent of the data myself, or approximately 2,500 instances, each instance being an observation contained in one line of an excel file. I received instruction on the three categories I chose and also a copy of the category definitions from Robinson. Because the original data lacks temporal attributes, I had to use the re-annotated data for drift analysis.

I spent approximately 12 hours re-coding this data, spaced over several days. To analyze drift in the coding process, I used only the 10% of the data that I personally re-coded, which is a smaller corpus than the original data.

Inter-Rater Reliability Quantification

I used Cohen's Kappa to quantify agreement. This is a departure from focusing on standard machine learning F measures that average misses and false alarms for the evaluation of classifier performance. Cohen's Kappa in ML relates to the training data. However, Cohen's Kappa takes class imbalance into account and is consistent with practice in behavioral research, and is in fact included in WEKA output.

Analysis Approach

Measures

As just noted, I used Kappa to quantify agreement. I also performed a qualitative analysis to check the similarities between the features used by the ML program and those used by humans. If the ML program used key words for categorization that are related to the concepts in the instructions, then that indicates it is using a similar set of criteria as the human annotators. For example, if the program used key words associated with “reassurance” to categorize a behavior as patient management, that would be similar to the original annotation criteria. The list of stop words to inform feature analysis appears in NLTK’s list of English Stopwords (n.d.).

Results

RQ1: What is the quantified inter-rater reliability between the human and machine annotators?

Table 2 shows the Kappa values as well as the standard ML F values and values for precision and recall assuming the original annotation as ground truth for each of the three variables when running the full data set annotated by the original annotator for all three of our chosen variables through WEKA. I found a Cohen's Kappa of 0.83 for information gathered from the patient (PT), 0.60 for system management behavior (SM), and 0.18 for behavior related to logistics (LOG). These variables represented good balance, medium balance, and very poor balance respectively.

Table 2. ML Values for Original Annotated Data (Full Corpus)

	Kappa	**F Statistic**	**Precision**	**Recall**
Patient (PT)	0.83	0.935	0.935	0.935
System Management (SM)	0.60	0.907 *0.91	0.907	0.913
Logistics (LOG)	0.18	0.92	0.913	0.934

Note. Values reported are weighted averages as calculated by WEKA

Note. *F statistic calculated for SM as a Harmonic mean of Precision and Recall is different from that provided by WEKA and hence has provided explicitly

RQ2: How do the machine learning categorization terms compare with the original class definitions?

The *decision tree for PT* prioritized terms like "asks", "listens", "feels", "confirms", "yes", and "pt". All of these terms are associated with the original definition for the category, which is information gathering from the patient themself. Terms like “asks” and “pt”, which is the abbreviation for patient used in the data, are particularly telling in this respect.

The *tree for SM* prioritized terms such as “tsheets”, “tsheet”, “documents”, “pages”, and “log”. All of these terms are associated with the definition for system management, which deals primarily with paperwork, administrative tasks, and other ways of dealing with the system.

The *LOG decision tree* prioritized terms such as “ride”, “needed”, “wants”, “wont” “isnt”, and “the”. The logistics category includes behaviors that maximize patient benefit, minimizing resources used, and finding the most appropriate treatment. Terms such as “ride”, “needed” and “wants” are connected to this definition.

The most important words in the decision tree for each of the three categories are different, indicating that there is a quantifiable, conceptual difference between the categories.

RQ3: Do I need to eliminate stop words (highly frequent, typically non-discriminators e.g., “the”) from the data set to suppress spurious feature selection for qualitative data classification?

While stop words such as “she” and “if” are present in the PT decision tree, they are not among the top terms the decision tree uses to determine categorization for the data. The SM decision tree also includes some stop words such as “who” and “for”, but like the PT tree they are not high priority for the algorithm to make decisions about categorization.

Stop words were present in the LOG tree, and higher in priority than in the other two trees. Terms like “isn't” and “the” appeared relatively high on the tree for the LOG category. These terms are stop words, not directly associated with the category definition. The fact that I found these stop words higher on the decision tree for the most poorly balanced variable suggests that the ML algorithm has a more difficult time not only simply categorizing, but also identifying meaningful features for poorly balanced variables.

RQ4: What is the effect of drift in the human annotation process?

I analyzed the re-categorized data to check for the impact of drift in human annotation over time. I did this by splitting the re-categorized data into two halves chronologically, one half being the first one categorized and the other being the second. I chose to use the most well-balanced variable (patient information gathering) to remove the complicating factor of poor class balance. I processed each of these halves through WEKA to determine if the ML algorithm could detect a difference between them, which would indicate that there was a difference in how the two halves were categorized by the human rater. The Kappa value I found for the first half of the data was 0.72, and for the second half 0.76, confirming that there is no difference between the two halves of the data.

The most important words in both decision trees were the words “asks”, “pt”, and “res”, three terms that indicate the doctor asking the patient a question. These terms also appeared high up on the decision tree calculated for the full PT corpus. These results show that in this re-annotated data set, drift was not a significant problem for the human annotator. However, it is possible that I didn’t see drift in this annotation because a re-annotation was done on a smaller subset of the data.

RQ5: What is the quantified inter-rater reliability between two human annotators?

Complete recategorized corpus. When I re-categorized the data myself, I calculated Kappa again between my categorization and the previous human categorization. I found values of 0.95 for PT, 0.91 for SM and 0.78 for the LOG category. The Kappa values for human annotated data are higher for the more poorly balanced variables than for the well balanced one, although the well balanced variable has a similar Kappa between WEKA and the human annotators. I also ran the re-annotated data through WEKA, in order to compare the results to the results of the original annotation. Table 3 shows these WEKA-calculated Kappa values as well as F values and values for precision and recall for each of the three re-categorized variables.

Table 3. Values for Re-Annotated Data (Reduced Corpus)

	Human Kappa	**WEKA (ML) Kappa**	**F Statistic**	**Precision**	**Recall**
Patient	0.95	0.77	0.904	0.904	0.904
System Management	0.91	0.51	0.92	0.926	0.931
Logistics	0.78	0.17	0.914	0.906	0.928

RQ6: How does class imbalance affect machine learning and human annotation?

As expected, the Kappa value for a well-balanced variable was best and degraded as the balance became poorer. After using SMOTE in R to re-balance the full LOG corpus, I ran the re-balanced corpus through WEKA again. This data set gave me a significantly improved Kappa over the original, with a Kappa of 0.93. Table 4 shows the WEKA-calculated Kappa, F, precision and recall values for the re-balanced data. Because the data was balanced, approximately half positive results of logistics behavior and half negative, the ML algorithm was able to categorize it with a much higher degree of success. I saw a definite benefit to Cohen's Kappa calculation in WEKA as a result of re-balancing the data which could result merely from a reduction in the chance penalty.

However, the decision tree changed, prioritizing terms such as "felt", "tsheet", "reception", "because", "sign" and "none". Some of these terms, such as "felt", "sign" or "none" could be conceptually linked to the category definition of logistics behavior. The term "felt" could be linked to what resources they need for their condition. The terms “sign” and “none” could be indicators of what it is that the patient needs as well. The prioritization of stop words in the decision trees differed between the different levels of balance, with stop words appearing at a higher priority in the most poorly balanced variable. Also, it seems that the algorithm also struggled to maintain the conceptual difference between the most poorly balanced variable and other variables, as the terms in the LOG tree are less conceptually connected to the category. The most worrisome of these terms is “tsheet”, which is most often associated with the doctor doing paperwork, and is therefore connected to system management rather than logistics. Unlike the decision tree for the unbalanced LOG data, this tree did not prioritize stop words as high up on the tree. However, it also differed from all the other decision trees in another way. This tree does not branch as much as the other trees, and continues looking at new terms for much longer before making a decision and starting a new branch. This may be a result of the re-balanced LOG data having a smaller and less varied vocabulary to draw from for the positive instances, as SMOTE re-balancing creates new positive instances by oversampling from the existing ones.

Table 4. ML Values for Data Balanced With SMOTE

	Kappa	**F Statistic**	**Precision**	**Recall**
Logistics	0.93	0.964	0.966	0.964

Discussion

The intention of this study was to examine the possibility of using machine learning algorithms as a tool for assisting inter-rater reliability by substituting for the second rater. In order to conclude that they are a useful tool, I must determine whether I can use them to get good inter-rater reliability measurements, and whether the algorithm's decision rules are comparable to the category definitions. When human raters perform inter-rater reliability annotation, we assume that the second rater is using the features identified in the coding scheme. However, it is possible that a human rater could be sensitive to spurious correlated features, just as an ML algorithm could be. Benefits of using ML algorithms to replace the second rater include reducing the workload for human researchers who have to annotate qualitative data, as well as creating ML models that can be run repeatedly on multiple data sets, which is a more extensive analysis than researchers usually perform in psychological studies

Evidence Favoring the Adoption of ML for Inter-Rater Reliability Annotation

Both quantitative and qualitative analyses favor the adoption of ML. The results of using ML algorithms to substitute for one human rater in inter-rater reliability measurements seem positive, given good class balance of the data in question. When the data is relatively well balanced, the ML algorithm is comparable to a human rater. I did not find evidence that it is any better. The fact that I found acceptable inter-rater reliability between machine and human raters for good and moderate class balance, as seen by the calculated Cohen's Kappa values, supports the use of ML for this purpose.

Qualitative (or conceptual) disagreement might have appeared as completely different criteria to categorize the data, or relevant but surprising features that the original categorization guidelines did not take into account. This could happen because Random Forest ML algorithms are not given instructions and generate criteria to sort the data without researcher input. Again, given reasonable class balance, this did not appear to be a problem. Stop words also did not become a problem except in cases of very poor class balance in the data. This suggests that as long as the data is well balanced (with at least 15% positive instances) researchers can expect that the ML algorithm will have good conceptual agreement with the original classification guidelines.

Limitations of ML for the Adoption of ML for Inter-Rater Reliability Annotation

I have identified class imbalance as a major limitation. The results of my final analysis showed that the ML algorithm resulted in bad inter-rater reliability scores when the variables were more poorly balanced, as predicted. In contrast, the human-human inter-rater reliability scores maintained relatively high values across different levels of balance. This suggests that the second human coder apprehended the context or nuances of the English language that were not accessible to the machine. Features that the ML algorithm used to classify the data tend to degrade with degraded balance. More poorly balanced data resulted in more stop words being labeled as important for the classification. Unfortunately, SMOTE gave mixed results as a re-balancing tool. On the positive side, the Kappa results were good. The re-balanced data set produced a decision tree that included more content words at higher levels of importance. However, the relationship of those words to the category definition was tenuous. Nevertheless, because I looked at Cohen's Kappa, rather than F values, I demonstrated limitations of the ML algorithm that are not apparent in F alone.

Study Limitations

Because this study was focusing on the viability of ML algorithms as a substitute for the second rater, I did not examine alternative WEKA parameters. Future researchers should look into different classification methods and different parameter settings in WEKA. The specific results may be dependent on the settings used, and could even improve. Because stop-words appeared to be problematic only in the case of the poorly balanced class, I did not examine potential improvements with their removal in the well balanced classes, though this could be checked.   In this initial study, I examined the three labels in separate classification exercises. A multi-class classification algorithm, considering all labels simultaneously, is an important next step.

The sensitivity of my results to class balance limits the use of machine learning in many applications. This is a known limitation in the Information Retrieval literature, and therefore not surprising. From my results, it is unclear whether SMOTE was an effective solution for restoring class balance. While the data that was re-balanced using SMOTE produced a good Kappa value when run through WEKA, it also produced unclear decision tree results. It may be worthwhile for future research to explore alternative methods for class re-balancing or smoothening, such as with GANS. Data rebalanced with SMOTE only replicates existing data to obtain improved class balance; GANS would attempt to replicate patterns in the existing data by generating data from similar distribution, possibly producing better results. In either case, any re-balancing intervention raises a methodological concern. Potentially, human coders should receive a re-balanced corpus, although the human-human agreement did not fall off as much with class imbalance as the human-machine agreement.   In the case of human-machine agreement, perhaps reporting should include both balanced and unbalanced results. In addition, alternative balance penalties in the agreement statistic, such as Gwet's AC1 merit consideration. A completely different approach to manage class imbalance is to be less reliant on pattern discovery by providing guidance to the algorithm in the form of human curated rules. This implies a hybrid ML-human coding procedure as a candidate for future research.

My analysis assumed that the directly available text provided the required features. This is suspect. Perhaps synonyms and lexical variants *(asks, questions)* or even more challenging n-grams *(wants to know)* are better represented at a higher level of abstraction than is directly available in the text. This is particularly concerning because it is the observer who chose the descriptive language that appears in the field notes, though likely favoring simpler, shorter alternatives. Two approaches to alternative representation counter this problem. First, the text might be pre-processed using external knowledge bases and syntactic parsing to convert the text to a more abstract representation of its features. Second, it may also be helpful to represent the data as a vector when processing it with ML, rather than as text as I did. This will also pull latent features out of the data, without appeal to external knowledge bases. Vector representation also has the added advantage of not depending on the format of the data. The trouble with the method is that it loses the semantics of the class labels. It would return a vector representing the content with unnamed parameters, rather than a decision tree with identifiable text features that can be compared to the original class definitions. This might be attempted after more favorable results with the basic, lexically based approach used here.

I also had issues with my data set structured as a continuous stream. Each individual chunk of text could represent multiple instances of multiple variables. I resolved this with binary classification, but this still clouds the results and could have an impact on the algorithm's ability to pick up on important features. Binary classification is more of a patch than an actual solution.

Generalizability to other independent data sets was not tested here, and typically is not required in order to publish subsequent analyses with coded qualitative data. From a practical perspective, it may not matter if an individual ML algorithm can not generalize to all data. Building new ones to fit new sets of data is cheap and relatively easy, although they do require very large training data sets on which to train.

Because I was unable to look at the full corpus to examine drift in the human rater, I was not able to gain a full picture of how drift affects human ratings over time. The amount of data I was able to analyze for drift simply was not enough to effectively detect drift in human annotation over time. Future research in this area would benefit from examining drift more closely. If ML algorithms can alleviate the impact of drift that would be an added benefit to their use, and if they are impacted negatively by drift that would need to be taken into account.

Conclusion

Because my results showed a significant agreement between the human raters as well as agreement between the human and ML algorithm in cases of well-balanced data, clearly the algorithm is a useful substitute for human annotators in some circumstances. I found that ML algorithms can be used to categorize data with a strong agreement with human annotators. I did not have issues related to the human raters disagreeing with one another, indicating that the original classification scheme for the data was reliable. In-depth analysis showed both good agreement between human and ML annotators and good levels of conceptual agreement between the algorithm and the classification rules. Because of this, I can conclude that this data and its classification scheme are reasonable for testing the ML algorithm for this purpose, and more generally, that the original coding scheme is reliable.

Along with the practical benefits of using ML as a tool, this research will help researchers by helping to establish the reliability of classification schemes. ML algorithms are a promising tool for qualitative researchers who often need to categorize large amounts of qualitative data and check the reliability of that categorization.

Appendix A. Example Data as Coded by Robinson (2011)

Res Action	Info Gathering
new pt; abdominal pain lady, loopy, saw something in her cup at church's, feels sick, MRDD (mentally retarded)	1
we go into the room	0
res shakes hands with the pt	1
res asks about any fever	1
res asks pt where her belly hurts	1
res asks if that could just be from throwing up	0
res asks how many times pt has thrown up	1
res asks if pt has been throwing up clear liquid	1
res asks about any belly surgery	1
res asks about appendix and gallbladder	1
res asks if she has taken anything	1
res asks if it helped	1
res asks when it started	1
res listens to pt's chest and belly	1
res feels pt's belly	1
res asks if pt's bowels have been ok	1
res asks what color it is	1
res asks how many times pt has had diarrhea	1
res asks who pt's doc is	1
res asks about any other problems	1
res asks pt if she's on meds for it	1
res asks about a hysterectomy	1
res says the labs are back and look ok, they're still waiting on one	0
res says they'll give some nausea meds (that dissolve on the tongue) and try to get food to	0

stay down	
res asks what pt's drink was	1
res asks pt if she feels hungry but just can't eat or keep it down	1
we leave the room; res says it seems visceral (just an emotional reaction to the sight in her drink)	0
res talks to the att; gives pt's history, says she looks good, moist lips, not dry/dehydrated	0
res says nothing to add lab wise, will do zophran and an oral challenge	0
res tells me with belly you always rule out a surgical issue but the pt is nontender, no localized pain, wants to eat, no fever	0

Appendix B. WEKA Settings

We used a j48 tree classifier to classify the data in WEKA. In pre-processing, I translated the text data through the stringtowordvector filter and set the category data as the class. I have included the list of steps I went through to process the data from the beginning.

1. Open terminal

2. cd ~/Desktop/WEKA_Analysis

3. Change the file input name and output name

4. python3 file_converter.py

5. start weka

6. open newly created file *_mod.csv

7. save the file as *.arff

8. Remove line number 3 that starts with @attribute to @attribute tweet string

9. Change the next @attribute label numeric to @attribute class {0,1}

10. Open weka

11. Open the new arff file

12. click "infogathering"

13. In filter.. click choose

14. in unsupervised--attribute--StringToWordVector

If you choose something besides StringToWordVector, it might change the results

15. click apply

16. click label

17. click edit on upper right hand corner

18. right click on the first attribute list (first column) and choose "attribute as class". click ok

19. got to "classify" tab

20. click choose

21. select trees--j48

    -Alternatives: Logistic Regression (Logistic). NaiveBayes.

    -For situations with insufficient data

22. click start

-23. Select Percentage split to run the model on the reserved test set

Appendix C. Pilot Data WEKA Output File

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     INFO GATHERING MASTER_mod-
weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-N0-
stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-
M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters "
[...]
Instances:    10314
Attributes:   1445
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

asks <= 0
|   listens <= 0
|   |   feels <= 0
|   |   |   looks <= 0
|   |   |   |   checks <= 0
|   |   |   |   |   BMP <= 0
|   |   |   |   |   |   notes <= 0
|   |   |   |   |   |   |   student <= 0
|   |   |   |   |   |   |   |   adds <= 0
|   |   |   |   |   |   |   |   |   new <= 0
|   |   |   |   |   |   |   |   |   |   nurse <= 0
|   |   |   |   |   |   |   |   |   |   |   CBC <= 0
|   |   |   |   |   |   |   |   |   |   |   |   sees <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   orders <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   res <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   att <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   phone <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   admitted <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   to <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   resident <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   you <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pain <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pt <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   or <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   not <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   the <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   a <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chart <= 0: 1 (107.0/21.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | chart > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | a > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | for <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | than <= 0: 0 (11.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | than > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | for > 0: 1 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | the > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | and <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | are <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | negative <= 0: 0 (99.0/13.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | negative > 0: 1 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | are > 0: 1 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | and > 0: 1 (6.0)
| | | | | | | | | | | | | | | | | | | | | | | | not > 0
| | | | | | | | | | | | | | | | | | | | | | | | | back <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | xray <= 0: 1 (4.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | xray > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | back > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | or > 0: 1 (10.0)
| | | | | | | | | | | | | | | | | | | | | | pt > 0
| | | | | | | | | | | | | | | | | | | | | | | now <= 0
| | | | | | | | | | | | | | | | | | | | | | | | be <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | tells <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | up <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | a <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | the <= 0: 1 (22.0/4.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | the > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | they <= 0: 0 (17.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | they > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | a > 0: 0 (9.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | up > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | tells > 0: 0 (4.0)
| | | | | | | | | | | | | | | | | | | | | | | | be > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | now > 0: 0 (4.0)
| | | | | | | | | | | | | | | | | | | | | pain > 0: 1 (18.0/1.0)
| | | | | | | | | | | | | | | | | | | | you > 0: 1 (15.0)
| | | | | | | | | | | | | | | | | | | resident > 0
| | | | | | | | | | | | | | | | | | | | in <= 0
| | | | | | | | | | | | | | | | | | | | | negative <= 0
| | | | | | | | | | | | | | | | | | | | | | pain <= 0
| | | | | | | | | | | | | | | | | | | | | | | patients <= 0: 0 (30.0/2.0)
| | | | | | | | | | | | | | | | | | | | | | | patients > 0: 1 (4.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | pain > 0: 1 (3.0/1.0)
| | | | | | | | | | | | | | | | | | | | | negative > 0: 1 (3.0/1.0)
| | | | | | | | | | | | | | | | | | | | in > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | to > 0
| | | | | | | | | | | | | | | | | | | pts <= 0
| | | | | | | | | | | | | | | | | | | | you <= 0: 0 (228.0/14.0)
| | | | | | | | | | | | | | | | | | | | you > 0: 1 (3.0/1.0)
| | | | | | | | | | | | | | | | | | | pts > 0: 1 (3.0/1.0)
| | | | | | | | | | | | | | | | | admitted > 0: 0 (24.0/1.0)
| | | | | | | | | | | | | | | | phone > 0: 0 (22.0)
| | | | | | | | | | | | | | | att > 0: 0 (193.0/18.0)
| | | | | | | | | | | | | | res > 0
| | | | | | | | | | | | | | | documents <= 0
| | | | | | | | | | | | | | | | prescription <= 0
| | | | | | | | | | | | | | | | | by <= 0
| | | | | | | | | | | | | | | | | | does <= 0
| | | | | | | | | | | | | | | | | | | phone <= 0
| | | | | | | | | | | | | | | | | | | | tsheets <= 0
| | | | | | | | | | | | | | | | | | | | | says <= 0
| | | | | | | | | | | | | | | | | | | | | | has <= 0
| | | | | | | | | | | | | | | | | | | | | | | writes <= 0
| | | | | | | | | | | | | | | | | | | | | | | | tells <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | pts <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | talks <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | with <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | in <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | at <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | confirms <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | test <= 0: 0
(340.0/65.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | test > 0: 1 (4.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | confirms > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | is <= 0: 1 (9.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | is > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | at > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | sheets <= 0: 1
(14.0/2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | sheets > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | in > 0: 0 (47.0/6.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | with > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | check <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | of <= 0: 0 (58.0/2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | of > 0: 1 (3.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | check > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | talks > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | is <= 0: 0 (27.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | is > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | att <= 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | att > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | pts > 0
| | | | | | | | | | | | | | | | | | | | | | | | | | about <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | himself <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | family <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | goes <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | be <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | cant <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | story <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | up <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | lifts <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | leg <= 0: 1
(50.0/9.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | leg > 0: 0
(2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | lifts > 0: 1
(2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | up > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | story > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | cant > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | be > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | goes > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | family > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | himself > 0: 0 (5.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | about > 0: 0 (6.0)
| | | | | | | | | | | | | | | | | | | | | | | | tells > 0: 0 (120.0/6.0)
| | | | | | | | | | | | | | | | | | | | | | | writes > 0: 0 (49.0)
| | | | | | | | | | | | | | | | | | | | | | has > 0
| | | | | | | | | | | | | | | | | | | | | | | pt <= 0: 0 (11.0)
| | | | | | | | | | | | | | | | | | | | | | | pt > 0
| | | | | | | | | | | | | | | | | | | | | | | | tells <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | up <= 0: 1 (51.0/7.0)
| | | | | | | | | | | | | | | | | | | | | | | | | up > 0: 0 (3.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | tells > 0: 0 (9.0)
| | | | | | | | | | | | | | | | | | | | | says > 0
| | | | | | | | | | | | | | | | | | | | | | high <= 0
| | | | | | | | | | | | | | | | | | | | | | | gets <= 0: 0 (2397.0/197.0)
| | | | | | | | | | | | | | | | | | | | | | | gets > 0
| | | | | | | | | | | | | | | | | | | | | | | | xray <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | maybe <= 0: 0 (31.0/3.0)
| | | | | | | | | | | | | | | | | | | | | | | | | maybe > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | | | | xray > 0: 1 (3.0)
| | | | | | | | | | | | | | | | | | | | | | high > 0
| | | | | | | | | | | | | | | | | | | | | | | negative <= 0: 0 (42.0/9.0)
| | | | | | | | | | | | | | | | | | | | | | | negative > 0: 1 (3.0)
| | | | | | | | | | | | | | | | | | | | tsheets > 0: 0 (151.0/3.0)
| | | | | | | | | | | | | | | | | | | phone > 0: 0 (89.0/1.0)
| | | | | | | | | | | | | | | | | | does > 0
| | | | | | | | | | | | | | | | | | | rectal <= 0
| | | | | | | | | | | | | | | | | | | | US <= 0
| | | | | | | | | | | | | | | | | | | | | has <= 0
| | | | | | | | | | | | | | | | | | | | | | foot <= 0: 0 (59.0/5.0)
| | | | | | | | | | | | | | | | | | | | | | foot > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | has > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | US > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | rectal > 0: 1 (4.0)
| | | | | | | | | | | | | | | | | by > 0: 0 (36.0)
| | | | | | | | | | | | | | | | prescription > 0: 0 (36.0)
| | | | | | | | | | | | | | | documents > 0: 0 (141.0/1.0)
| | | | | | | | | | | | | orders > 0
| | | | | | | | | | | | | | xray <= 0
| | | | | | | | | | | | | | | CT <= 0
| | | | | | | | | | | | | | | | does <= 0
| | | | | | | | | | | | | | | | | for <= 0
| | | | | | | | | | | | | | | | | | pt <= 0
| | | | | | | | | | | | | | | | | | | LFTs <= 0
| | | | | | | | | | | | | | | | | | | | UA <= 0
| | | | | | | | | | | | | | | | | | | | | US <= 0
| | | | | | | | | | | | | | | | | | | | | | cardiac <= 0
| | | | | | | | | | | | | | | | | | | | | | | enzymes <= 0: 0 (52.0/9.0)
| | | | | | | | | | | | | | | | | | | | | | | enzymes > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | | cardiac > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | | US > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | | UA > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | | LFTs > 0: 1 (2.0)
| | | | | | | | | | | | | | | | | | pt > 0: 1 (6.0/2.0)
| | | | | | | | | | | | | | | | | for > 0: 0 (19.0/1.0)
| | | | | | | | | | | | | | | | does > 0: 1 (4.0)
| | | | | | | | | | | | | | | CT > 0: 1 (10.0/1.0)
| | | | | | | | | | | | | | xray > 0: 1 (18.0/2.0)
| | | | | | | | | | | | sees > 0
| | | | | | | | | | | | | add <= 0
| | | | | | | | | | | | | | big <= 0
| | | | | | | | | | | | | | | here <= 0
| | | | | | | | | | | | | | | | in <= 0: 1 (31.0/10.0)
| | | | | | | | | | | | | | | | in > 0: 0 (2.0)
| | | | | | | | | | | | | | | here > 0: 0 (2.0)
| | | | | | | | | | | | | | big > 0: 0 (2.0)
| | | | | | | | | | | | | add > 0: 0 (2.0)
| | | | | | | | | | | CBC > 0
| | | | | | | | | | | | out <= 0
| | | | | | | | | | | | | pt <= 0
| | | | | | | | | | | | | | was <= 0
| | | | | | | | | | | | | | | res <= 0
| | | | | | | | | | | | | | | | low <= 0: 0 (2.0)
| | | | | | | | | | | | | | | | low > 0: 1 (2.0)
| | | | | | | | | | | | | | | res > 0: 1 (12.0)
| | | | | | | | | | | | | | was > 0: 0 (3.0)
| | | | | | | | | | | | | pt > 0: 0 (3.0)
| | | | | | | | | | | | out > 0: 0 (3.0)
| | | | | | | | | | nurse > 0
| | | | | | | | | | | says <= 0
| | | | | | | | | | | | blood <= 0
| | | | | | | | | | | | | hands <= 0
| | | | | | | | | | | | | | much <= 0: 0 (71.0/9.0)
| | | | | | | | | | | | | | much > 0: 1 (2.0)
| | | | | | | | | | | | | hands > 0: 1 (3.0/1.0)
| | | | | | | | | | | | blood > 0
| | | | | | | | | | | | | orders <= 0: 0 (3.0/1.0)
| | | | | | | | | | | | | orders > 0: 1 (2.0)
| | | | | | | | | | | says > 0
| | | | | | | | | | | | doc <= 0
| | | | | | | | | | | | | the <= 0
| | | | | | | | | | | | | | a <= 0: 1 (42.0)
| | | | | | | | | | | | | | a > 0
| | | | | | | | | | | | | | | pt <= 0: 0 (6.0/1.0)
| | | | | | | | | | | | | | | pt > 0: 1 (10.0)
| | | | | | | | | | | | | the > 0
| | | | | | | | | | | | | | about <= 0
| | | | | | | | | | | | | | | cath <= 0
| | | | | | | | | | | | | | | | IV <= 0
| | | | | | | | | | | | | | | | | give <= 0
| | | | | | | | | | | | | | | | | | he <= 0
| | | | | | | | | | | | | | | | | | | tell <= 0
| | | | | | | | | | | | | | | | | | | | tells <= 0
| | | | | | | | | | | | | | | | | | | | | too <= 0
| | | | | | | | | | | | | | | | | | | | | | bathroom <= 0
| | | | | | | | | | | | | | | | | | | | | | | up <= 0: 1 (39.0/3.0)
| | | | | | | | | | | | | | | | | | | | | | | up > 0: 0 (3.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | bathroom > 0: 0 (2.0)
| | | | | | | | | | | | | | | | | | | | | too > 0: 0 (5.0/1.0)
| | | | | | | | | | | | | | | | | | | | tells > 0: 0 (6.0/2.0)
| | | | | | | | | | | | | | | | | | | tell > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | | he > 0: 0 (3.0)
| | | | | | | | | | | | | | | | | give > 0: 0 (3.0)
| | | | | | | | | | | | | | | | IV > 0: 0 (3.0)
| | | | | | | | | | | | | | | cath > 0: 1 (5.0)
| | | | | | | | | | | | | | about > 0: 0 (7.0)
| | | | | | | | | | | | doc > 0: 0 (11.0/1.0)
| | | | | | | | | new > 0
|  |  |  |  |  |  |  |  |  |  to <= 0
|  |  |  |  |  |  |  |  |  |  |  pain <= 0
|  |  |  |  |  |  |  |  |  |  |  |  afib <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  with <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  because <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  res <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  a <= 0: 1 (23.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  a > 0: 0 (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  res > 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  EKG <= 0: 0 (31.0/9.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  EKG > 0: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  because > 0: 1 (5.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  with > 0: 1 (12.0)
|  |  |  |  |  |  |  |  |  |  |  |  afib > 0: 0 (4.0)
|  |  |  |  |  |  |  |  |  |  |  pain > 0: 1 (20.0/1.0)
|  |  |  |  |  |  |  |  |  |  to > 0
|  |  |  |  |  |  |  |  |  |  |  seen <= 0: 0 (23.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  seen > 0: 1 (2.0)
|  |  |  |  |  |  |  |  adds > 0
|  |  |  |  |  |  |  |  |  on <= 0
|  |  |  |  |  |  |  |  |  |  atavan <= 0
|  |  |  |  |  |  |  |  |  |  |  nitro <= 0
|  |  |  |  |  |  |  |  |  |  |  |  problem <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  for <= 0: 1 (30.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  for > 0: 0 (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  problem > 0: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  nitro > 0: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  atavan > 0: 0 (2.0)
|  |  |  |  |  |  |  |  |  on > 0: 0 (4.0)
|  |  |  |  |  |  |  student > 0
|  |  |  |  |  |  |  |  had <= 0
|  |  |  |  |  |  |  |  |  if <= 0
|  |  |  |  |  |  |  |  |  |  new <= 0
|  |  |  |  |  |  |  |  |  |  |  says <= 0
|  |  |  |  |  |  |  |  |  |  |  |  at <= 0: 0 (24.0/5.0)
|  |  |  |  |  |  |  |  |  |  |  |  at > 0: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  says > 0
|  |  |  |  |  |  |  |  |  |  |  |  home <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  then <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  though <= 0: 1 (45.0/6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  though > 0: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  then > 0: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  home > 0: 0 (3.0)
|  |  |  |  |  |  |  |  |  |  new > 0: 1 (4.0)
|  |  |  |  |  |  |  |  |  if > 0: 0 (7.0/1.0)
|  |  |  |  |  |  |  |  had > 0: 1 (15.0)
|   |   |   |   |   |   notes > 0
|   |   |   |   |   |   |   but <= 0
|   |   |   |   |   |   |   |   xray <= 0
|   |   |   |   |   |   |   |   |   for <= 0
|   |   |   |   |   |   |   |   |   |   do <= 0
|   |   |   |   |   |   |   |   |   |   |   his <= 0: 1 (48.0/5.0)
|   |   |   |   |   |   |   |   |   |   |   his > 0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   do > 0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   |   for > 0
|   |   |   |   |   |   |   |   |   |   has <= 0: 0 (4.0)
|   |   |   |   |   |   |   |   |   |   has > 0: 1 (3.0)
|   |   |   |   |   |   |   |   xray > 0: 0 (4.0)
|   |   |   |   |   |   |   but > 0: 0 (5.0)
|   |   |   |   |   BMP > 0
|   |   |   |   |   |   ddimer <= 0
|   |   |   |   |   |   |   att <= 0
|   |   |   |   |   |   |   |   normal <= 0
|   |   |   |   |   |   |   |   |   he <= 0: 1 (34.0/3.0)
|   |   |   |   |   |   |   |   |   he > 0
|   |   |   |   |   |   |   |   |   |   the <= 0: 0 (2.0)
|   |   |   |   |   |   |   |   |   |   the > 0: 1 (2.0)
|   |   |   |   |   |   |   |   normal > 0: 0 (2.0)
|   |   |   |   |   |   |   att > 0: 0 (3.0/1.0)
|   |   |   |   |   |   ddimer > 0: 0 (5.0/1.0)
|   |   |   |   checks > 0
|   |   |   |   |   on <= 0
|   |   |   |   |   |   to <= 0
|   |   |   |   |   |   |   back <= 0: 1 (147.0/20.0)
|   |   |   |   |   |   |   back > 0
|   |   |   |   |   |   |   |   a <= 0: 0 (7.0/2.0)
|   |   |   |   |   |   |   |   a > 0: 1 (2.0)
|   |   |   |   |   |   to > 0
|   |   |   |   |   |   |   give <= 0
|   |   |   |   |   |   |   |   blood <= 0
|   |   |   |   |   |   |   |   |   see <= 0: 0 (12.0/1.0)
|   |   |   |   |   |   |   |   |   see > 0
|   |   |   |   |   |   |   |   |   |   and <= 0: 1 (4.0)
|   |   |   |   |   |   |   |   |   |   and > 0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   blood > 0: 1 (2.0)
|   |   |   |   |   |   |   give > 0: 1 (5.0)
|   |   |   |   |   on > 0
|   |   |   |   |   |   med <= 0
|   |   |   |   |   |   |   meds <= 0
|   |   |   |   |   |   |   |   test <= 0
|   |   |   |   |   |   |   |   |   and <= 0
|   |   |   |   |   |   |   |   |   |   because <= 0
|   |   |   |   |   |   |   |   |   |   |   if <= 0: 0 (25.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   if > 0: 1 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   because > 0: 1 (3.0/1.0)
|   |   |   |   |   |   |   |   |   and > 0: 1 (3.0/1.0)
|   |   |   |   |   |   |   |   test > 0: 1 (2.0)
|   |   |   |   |   |   |   meds > 0: 1 (2.0)
|   |   |   |   |   |   med > 0: 1 (3.0)
|   |   |   looks > 0
|   |   |   |   at <= 0
|   |   |   |   |   eyes <= 0
|   |   |   |   |   |   in <= 0
|   |   |   |   |   |   |   up <= 0
|   |   |   |   |   |   |   |   tells <= 0
|   |   |   |   |   |   |   |   |   pretty <= 0
|   |   |   |   |   |   |   |   |   |   to <= 0
|   |   |   |   |   |   |   |   |   |   |   blood <= 0
|   |   |   |   |   |   |   |   |   |   |   |   back <= 0: 0 (68.0/16.0)
|   |   |   |   |   |   |   |   |   |   |   |   back > 0: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   blood > 0: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   to > 0: 0 (21.0/2.0)
|   |   |   |   |   |   |   |   |   pretty > 0
|   |   |   |   |   |   |   |   |   |   the <= 0: 0 (2.0)
|   |   |   |   |   |   |   |   |   |   the > 0: 1 (3.0)
|   |   |   |   |   |   |   |   tells > 0: 0 (12.0)
|   |   |   |   |   |   |   up > 0
|   |   |   |   |   |   |   |   be <= 0
|   |   |   |   |   |   |   |   |   for <= 0: 1 (14.0/2.0)
|   |   |   |   |   |   |   |   |   for > 0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   be > 0: 0 (3.0)
|   |   |   |   |   |   in > 0
|   |   |   |   |   |   |   for <= 0
|   |   |   |   |   |   |   |   says <= 0
|   |   |   |   |   |   |   |   |   res <= 0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   |   res > 0: 1 (43.0/2.0)
|   |   |   |   |   |   |   |   says > 0
|   |   |   |   |   |   |   |   |   cant <= 0
|   |   |   |   |   |   |   |   |   |   too <= 0: 0 (8.0/1.0)
|   |   |   |   |   |   |   |   |   |   too > 0: 1 (2.0)
|   |   |   |   |   |   |   |   |   cant > 0: 1 (2.0)
|   |   |   |   |   |   |   for > 0
|   |   |   |   |   |   |   |   reference <= 0: 0 (8.0/1.0)
|   |   |   |   |   |   |   |   reference > 0: 1 (2.0)
|   |   |   |   |   eyes > 0: 1 (25.0)
|   |   |   |   at > 0
|   |   |   |   |   tsheets <= 0
|   |   |   |   |   |   to <= 0
|  |  |  |  |  |  |  and <= 0
|  |  |  |  |  |  |  |  EKG <= 0
|  |  |  |  |  |  |  |  |  chart <= 0
|  |  |  |  |  |  |  |  |  |  orders <= 0: 1 (344.0/19.0)
|  |  |  |  |  |  |  |  |  |  orders > 0: 0 (4.0/1.0)
|  |  |  |  |  |  |  |  |  chart > 0: 0 (8.0/2.0)
|  |  |  |  |  |  |  |  EKG > 0: 1 (33.0)
|  |  |  |  |  |  |  and > 0
|  |  |  |  |  |  |  |  tsheet <= 0: 1 (89.0/6.0)
|  |  |  |  |  |  |  |  tsheet > 0: 0 (7.0)
|  |  |  |  |  |  to > 0
|  |  |  |  |  |  |  document <= 0
|  |  |  |  |  |  |  |  res <= 0: 0 (4.0/1.0)
|  |  |  |  |  |  |  |  res > 0
|  |  |  |  |  |  |  |  |  take <= 0
|  |  |  |  |  |  |  |  |  |  I <= 0: 1 (51.0/5.0)
|  |  |  |  |  |  |  |  |  |  I > 0: 0 (3.0/1.0)
|  |  |  |  |  |  |  |  |  take > 0: 0 (4.0/1.0)
|  |  |  |  |  |  |  document > 0: 0 (4.0/1.0)
|  |  |  |  |  tsheets > 0
|  |  |  |  |  |  takes <= 0
|  |  |  |  |  |  |  history <= 0: 0 (22.0/3.0)
|  |  |  |  |  |  |  history > 0: 1 (2.0)
|  |  |  |  |  |  takes > 0: 1 (3.0)
|  |  feels > 0
|  |  |  att <= 0
|  |  |  |  says <= 0: 1 (186.0/1.0)
|  |  |  |  says > 0
|  |  |  |  |  to <= 0
|  |  |  |  |  |  pts <= 0: 0 (5.0/1.0)
|  |  |  |  |  |  pts > 0: 1 (4.0)
|  |  |  |  |  to > 0: 0 (5.0)
|  |  |  att > 0: 0 (12.0/2.0)
|  listens > 0
|  |  res <= 0
|  |  |  att <= 0: 1 (8.0)
|  |  |  att > 0: 0 (2.0)
|  |  res > 0: 1 (179.0)
asks > 0
|  att <= 0
|  |  questions <= 0
|  |  |  wants <= 0
|  |  |  |  nurse <= 0
|  |  |  |  |  they <= 0
|  |  |  |  |  |  says <= 0
|  |  |  |  |  |  |  got <= 0
| | | | | | | | can <= 0
| | | | | | | | | meds <= 0: 1 (2623.0/92.0)
| | | | | | | | | meds > 0
| | | | | | | | | | needs <= 0: 1 (76.0/2.0)
| | | | | | | | | | needs > 0: 0 (6.0/1.0)
| | | | | | | | can > 0
| | | | | | | | | ride <= 0
| | | | | | | | | | pt <= 0: 0 (4.0/1.0)
| | | | | | | | | | pt > 0: 1 (37.0/3.0)
| | | | | | | | | ride > 0: 0 (3.0)
| | | | | | | got > 0
| | | | | | | | yet <= 0
| | | | | | | | | already <= 0: 1 (40.0/9.0)
| | | | | | | | | already > 0: 0 (2.0)
| | | | | | | | yet > 0: 0 (3.0)
| | | | | | says > 0
| | | | | | | has <= 0
| | | | | | | | another <= 0
| | | | | | | | | needs <= 0
| | | | | | | | | | them <= 0
| | | | | | | | | | | at <= 0
| | | | | | | | | | | | doc <= 0
| | | | | | | | | | | | | probably <= 0
| | | | | | | | | | | | | | but <= 0
| | | | | | | | | | | | | | | what <= 0
| | | | | | | | | | | | | | | | shes <= 0
| | | | | | | | | | | | | | | | | a <= 0: 1 (32.0/1.0)
| | | | | | | | | | | | | | | | | a > 0
| | | | | | | | | | | | | | | | | | in <= 0
| | | | | | | | | | | | | | | | | | | if <= 0: 0 (3.0/1.0)
| | | | | | | | | | | | | | | | | | | if > 0: 1 (11.0/2.0)
| | | | | | | | | | | | | | | | | | in > 0: 0 (2.0)
| | | | | | | | | | | | | | | | shes > 0: 0 (3.0/1.0)
| | | | | | | | | | | | | | | what > 0: 0 (5.0/2.0)
| | | | | | | | | | | | | | but > 0
| | | | | | | | | | | | | | | the <= 0: 1 (3.0)
| | | | | | | | | | | | | | | the > 0: 0 (5.0)
| | | | | | | | | | | | | probably > 0: 0 (3.0)
| | | | | | | | | | | | doc > 0
| | | | | | | | | | | | | pts <= 0: 0 (5.0)
| | | | | | | | | | | | | pts > 0: 1 (3.0/1.0)
| | | | | | | | | | | at > 0: 1 (4.0)
| | | | | | | | | | them > 0: 0 (4.0)
| | | | | | | | | needs > 0: 0 (4.0)
| | | | | | | | another > 0: 0 (5.0)
| | | | | | | has > 0: 1 (16.0/1.0)
|  |  |  |  |  they > 0
|  |  |  |  |  |  blood <= 0
|  |  |  |  |  |  |  ok <= 0
|  |  |  |  |  |  |  |  yet <= 0
|  |  |  |  |  |  |  |  |  want <= 0
|  |  |  |  |  |  |  |  |  |  get <= 0
|  |  |  |  |  |  |  |  |  |  |  do <= 0
|  |  |  |  |  |  |  |  |  |  |  |  says <= 0: 1 (37.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  |  says > 0
|  |  |  |  |  |  |  |  |  |  |  |  |  home <= 0: 1 (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  home > 0: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  do > 0: 0 (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  get > 0: 0 (3.0)
|  |  |  |  |  |  |  |  |  want > 0: 0 (7.0/1.0)
|  |  |  |  |  |  |  |  yet > 0: 0 (4.0)
|  |  |  |  |  |  |  ok > 0: 0 (4.0)
|  |  |  |  |  |  blood > 0: 0 (6.0)
|  |  |  |  nurse > 0
|  |  |  |  |  an <= 0
|  |  |  |  |  |  comes <= 0
|  |  |  |  |  |  |  sure <= 0
|  |  |  |  |  |  |  |  meds <= 0
|  |  |  |  |  |  |  |  |  want <= 0
|  |  |  |  |  |  |  |  |  |  now <= 0
|  |  |  |  |  |  |  |  |  |  |  been <= 0
|  |  |  |  |  |  |  |  |  |  |  |  bp <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  nitro <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  to <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  about <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  says <= 0: 0 (7.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  says > 0: 1 (12.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  about > 0: 0 (9.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  to > 0: 0 (14.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  nitro > 0: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  bp > 0: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  been > 0: 1 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  now > 0: 1 (4.0/1.0)
|  |  |  |  |  |  |  |  |  want > 0: 0 (3.0)
|  |  |  |  |  |  |  |  meds > 0: 1 (9.0/2.0)
|  |  |  |  |  |  |  sure > 0: 0 (5.0)
|  |  |  |  |  |  comes > 0: 1 (5.0)
|  |  |  |  |  an > 0: 1 (5.0)
|  |  |  wants > 0
|  |  |  |  nurse <= 0: 0 (36.0/3.0)
|  |  |  |  nurse > 0
|  |  |  |  |  orders <= 0
|   |   |   |   |   |   see <= 0: 1 (4.0)
|   |   |   |   |   |   see > 0: 0 (2.0)
|   |   |   |   |   orders > 0: 0 (3.0)
|   |   questions > 0: 0 (36.0/2.0)
|   att > 0: 0 (248.0/15.0)

Number of Leaves  : 268

Size of the tree :        535


Time taken to build model: 120.74 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9066               87.8999 %
Incorrectly Classified Instances      1248               12.1001 %
Kappa statistic                          0.7579
Mean absolute error                      0.1697
Root mean squared error                  0.3217
Relative absolute error                 33.9505 %
Root relative squared error             64.3479 %
Total Number of Instances            10314

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
              0.898    0.140    0.867      0.898    0.882      0.758    0.907     0.862    0
              0.860    0.102    0.892      0.860    0.876      0.758    0.907     0.887    1
Weighted Avg.    0.879    0.121    0.879      0.879    0.879      0.758    0.907     0.874

=== Confusion Matrix ===

    a    b   <-- classified as
 4674  533 |    a = 0
  715 4392 |    b = 1
```

References

262588213843476. (n.d.). *NLTK's list of ENGLISH STOPWORDS*. https://gist.github.com/sebleier/554280.

Auerbach, C. F., & Silverstein, L. B. (2003). *Qualitative data : an introduction to coding and analysis*. New York : New York University Press, c2003. Retrieved from http://ezproxy.libraries.wright.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat01902a&AN=wsu.b2278615&site=eds-live

Bhatt, S., Chen, K., Shalin, V. L., Sheth, A. P., & Minnery, B. (2019). *Who Should Be the Captain This Week? Leveraging Inferred Diversity-Enhanced Crowd Wisdom for a Fantasy Premier League Captain Prediction.*

Brooks, C., Amundson, K., Greer, J.: Detecting Significant Events in Lecture Video using Supervised Machine Learning. In: International Conference on Artificial Intelligence in Education (AIED), Brighton, UK (2009)

Burns, M. K. (2014). How to establish interrater reliability. *Nursing,44*(10), 56-58. doi:10.1097/01.nurse.0000453705.41413.c6

Creswell, J. W. (2007). *Qualitative inquiry and research design: choosing among five approaches* (2nd ed.). Sage.

Davidson, E., Edwards, R., Jamieson, L., & Weller, S. (2018). Big data, qualitative style: A breadth-and-depth method for working with large amounts of secondary qualitative data. *Quality & Quantity,53*(1), 363-376. doi:10.1007/s11135-018-0757-y

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. *2nd ed*. New York: Wiley.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1), 29-48.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1), 10-18.

Kang, O., & Johnson, D. O. (2015). Comparison of Inter-rater Reliability of Human and Computer Prosodic Annotation Using Brazil's Prosody Model. English Linguistics Research, 4(4), p58.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., . . . Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies,48*(6), 661-671. doi:10.1016/j.ijnurstu.2011.01.016

Kursuncu, U., Gaur, M., Castillo, C., Alambo, A., Thirunarayan, K., Shalin, V., Achilov, D., Arpinar, I. B., & Sheth, A. (2019). *Modeling Islamist Extremist Communications on Social Media using Contextual Dimensions: Religion, Ideology, and Hate*.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Maciejewski, T., & Stefanowski, J. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. *2011 IEEE Symposium on Computational Intelligence & Data Mining (CIDM)*, 104.

Pandey, R., Purohit, H., Castillo, C., & Shalin, V. L. (in review). *Modeling and Mitigating Human Annotation Errors to Design Efficient Stream Processing Systems with Human-in-the-loop Machine Learning*

Purohit, H., Hampton, A., Bhatt, S., Shalin, V. L., Sheth, A. P., Flach, J. M. (2014). Identifying Seekers and Suppliers in Social Media Communities to Support Crisis Coordination. *Comput Supported Coop Work* 23, 513–545. https://doi.org/10.1007/s10606-014-9209-y

Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013). What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination. *Computers in Human Behavior*, *29*(6), 2438–2447.

Robinson, F. E. (2011). *The Role of Deliberate Behavior in Expert Performance: The Acquisition of Information Gathering Strategy in the Context of Emergency Medicine* (Unpublished Master's thesis). Wright State University, Ohio.

Shankle WR, Mani S, Dick MB, Pazzani MJ. Simple models for estimating dementia severity using machine learning. Stud Health Technol Inform. 1998;52(pt 1):472-476.

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. Retrieved from http://ezproxy.libraries.wright.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=mnh&AN=15733050&site=eds-live

Williams, J.A., Weakley, A., Cook, D.J., Schmitter-Edgecombe, M.: Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In: Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)