Using Machine Learning to Predict Chat Difficulty

Jeremy Walker and Jason Coleman

This study aims to evaluate the effectiveness and potential utility of using machine learning and natural language processing techniques to develop models that can reliably predict the relative difficulty of incoming chat reference questions. Using a relatively large sample size of chat transcripts (N = 15,690), an empirical experimental design was used to test and evaluate 640 unique models. Results showed the predictive power of observed modeling processes to be highly statistically significant. These findings have implications for how library service managers may seek to develop and refine reference services using advanced analytical methods.

Introduction

Academic libraries face persistent challenges and open questions with regard to how they can best support and manage virtual reference services (VRS), sometimes referred to as "chat reference." Academic libraries are experiencing increasingly tight budgets, limited human resources, and, as a consequence of the COVID-19 pandemic, a potential sea-change in how and where librarians perform their work. Consequently, library service managers will need to continue to adapt and evolve in their approach to managing and delivering reference services. Many managers are looking increasingly to technology to support remote operations. Zoom chat services, IM chat, appointment scheduling, and real-time monitoring of space capacity have been receiving abundant attention. We suggest that this is also a perfect time to examine possible applications of machine learning and advanced text analysis.

Building upon previous work,[1] the authors organized a research project intended to explore if machine learning and natural language processing (NLP) methods can be used to develop models that effectively predict the relative *difficulty* of VRS service interactions based strictly on the opening questions and initial statements provided by patrons. The research project's scope was strictly focused on the structure and evaluation of predictive models.

Although beyond the scope of this paper, our research may have implications for future developments in VRS systems and managing patron services. Ideally, if any formulation of the models covered by this research are implemented in practice, library service managers may be able to effectively triage incoming inquiries to librarians, staff, and student-employees according to their respective skills, training, and duties. This has the potential to drastically improve librarian workloads by redirecting simple, directional, and rote questions away from highly skilled librarians and toward student employees, library interns, or other library employees.

Jeremy Walker is Data Services Librarian at Georgia State University; email: jwalker184@gsu.edu. Jason Coleman is Head, Library User Services, at Kansas State University; email: coleman@k-state.edu.

Literature Review

The vast majority of recent research concerning VRS transcripts has been inferential, qualitative, and small-scale in nature. A large body of research has focused on qualitative reviews and assessment of VRS operators' behavior with respect to standards, service quality, and operators' skills.[2] Other research has focused on qualitative reviews and assessment of VRS users' behavior during chat sessions with an emphasis on the types of questions users have and users' perceptions of library resources.[3] The qualitative and inferential nature of much of the research involving VRS, chat reference in particular, is ubiquitous in the literature.[4]

One particularly relevant vein of research pertaining to reference services is the evaluation and analysis of discrete categories of VRS operators. "Librarians," "Staff," "Student Assistants," and related synonyms are often used to categorize different types of VRS operators with respect to varying perceptions of labor costs, skills, and job duties with respect to libraries' operations. A relevant focal point in the literature is concerned with the ability of library staff, including VRS operators, to refer patrons to more qualified librarians and staff quickly and efficiently.[5]

Focusing on student employees, Bravender et al. found that the majority of analyzed VRS interactions did not require the skills of full-time librarians and that student operators were fully capable of providing high-quality service to users.[6] Most recently, the research from Radniecki and Winterman shows that the potential for student employees to provide high-quality service is not limited to routine and directional questions, but also more advanced and niche services.[7] Research also indicates that, while student employees are capable of providing excellent service, they are less skilled at referring patrons with challenging inquiries.[8] Critically, the combination of these insights implies that there is value in developing systems that can automatically triage incoming VRS inquiries to appropriate VRS operators, regardless of how they are segmented, to enhance the effectiveness of library service operations. Building on these implications, the authors of the current study sought to evaluate the utility of quantitative models for the potential purpose of triaging incoming VRS interactions based on the assumption that certain types of library employees could be designated to answer certain types of VRS questions.

Although there is robust research related to VRS and chat reference, the literature involving the use of "machine learning" and quantitative modeling of VRS chat transcripts is relatively sparse. In general, as with many other fields, "machine learning" and "artificial intelligence" continue to be popular and exciting areas of research and development in libraries. In recent years, repositories specifically designed to host information about "AI" projects have appeared, and many machine learning projects have been conducted in libraries, largely in the context of text mining and search.[9] Some applied research and development has been done with respect to multiple independent chatbot projects designed specifically to answer VRS inquiries that are rote or predictable in nature.[10] The relatively recent development of these projects indicates an increasing need for automated systems designed to enhance library service operations. However, information about these chatbot systems' performance is sparse, and only one project reported achieving approximately 50 percent "accuracy" on limited question-answering tasks.[11]

Exempting prior work conducted by the authors,[12] very little analysis and research appears to have been conducted on VRS transcripts at a large scale using quantitative methods. Kohler provides one of the only known examples of analysis using empirical, algorithmic methods for modeling and deriving insights from VRS transcripts pertaining to topic-modeling, sentiment analysis, and assigning difficulty ratings.[13] Of particular note, Kohler's research provides concrete examples of successfully using latent variable algorithms (examples: NMF, LDA, LSA) to extract and identify "topics" that manifest in chat transcripts and how these latent variables can be reliably mapped to READ Scale ratings; arguably the most widespread measure of "difficulty" for library-patron inquiries.[14] Unfortunately, since Kohler does not clearly articulate exactly how individual VRS transcripts are mapped to the READ Scale or provide any aggregated metrics of predictive accuracy, it is not possible to benchmark future research against Kohler's work.[15]

Since the research goals for the current study were not focused on topic-modeling, other methods were identified for the purposes of processing, computing, and modeling VRS transcripts. While the literature surrounding natural-language processing techniques and methods is immense and filled with minor variations and improvements upon established and foundational methods, only a narrow slice of the literature is emphasized here. The first identified method centers on eliminating infrequent and low-value words using *TF-IDF* metrics as described by Weiss et al.[16] Second, for the purposes of processing and quantifying sentences, Mikolov et al. introduced the Doc2Vec algorithm as a novel approach for converting documents of unequal length to fixed-length vectors.[17] This approach to quantifying and representing documents has gained traction and has been applied to research in clinical, genetic, and news-journalism fields.[18] Lau and Baldwin provide specific insights into useful practices for implementing the Doc2Vec algorithm.[19] Last, research has also shown that the incorporation of human-defined ontologies in the form of tags and labels provides much-needed structure to text analysis tasks.[20] This is reinforced by results derived from prior work indicating that the incorporation of ontology and domain-knowledge into predictive models focusing on VRS transcripts has a clear, positive, and statistically significant impact on overall model performance.[21] Taken as a whole, this suite of methods represents the foundation of the modeling processes implemented by the authors in prior work and the experimental design laid out in this paper.[22]

Methods

This study was conducted using a dataset collected at Kansas State University Libraries (KSUL). At KSUL, a combination of librarian faculty, specialists, staff, and students provide varying levels of patron services referred to as "Ask-A-Librarian." As a subset of these service operations, KSUL operates a chat reference service using the LibraryH3lp software embedded throughout KSUL's webpages.

The raw dataset used in this study consists of 15,690 individual samples collected between August 2013 and March 2018. Each individual sample represents a unique VRS chat transcript and accompanying metadata. This data was and continues to be created as individual VRS operators record chat transcripts in KSUL's reference tracking system (LibInsight) at the conclusion of each individual interaction. When logging submissions in LibInsight, VRS operators at KSUL are expected to record additional information such as *Question Type, READ Scale,* and *Time Spent* in addition to the raw text of the entire chat transcript. By default, all VRS interactions between users and operators are anonymous, and users have the reasonable expectation of anonymity and confidentiality with respect to using KSUL's VRS system. In the event that personally identifying information is present in an individual transcript, operators are expected to manually redact such information when recording data in LibInsight.

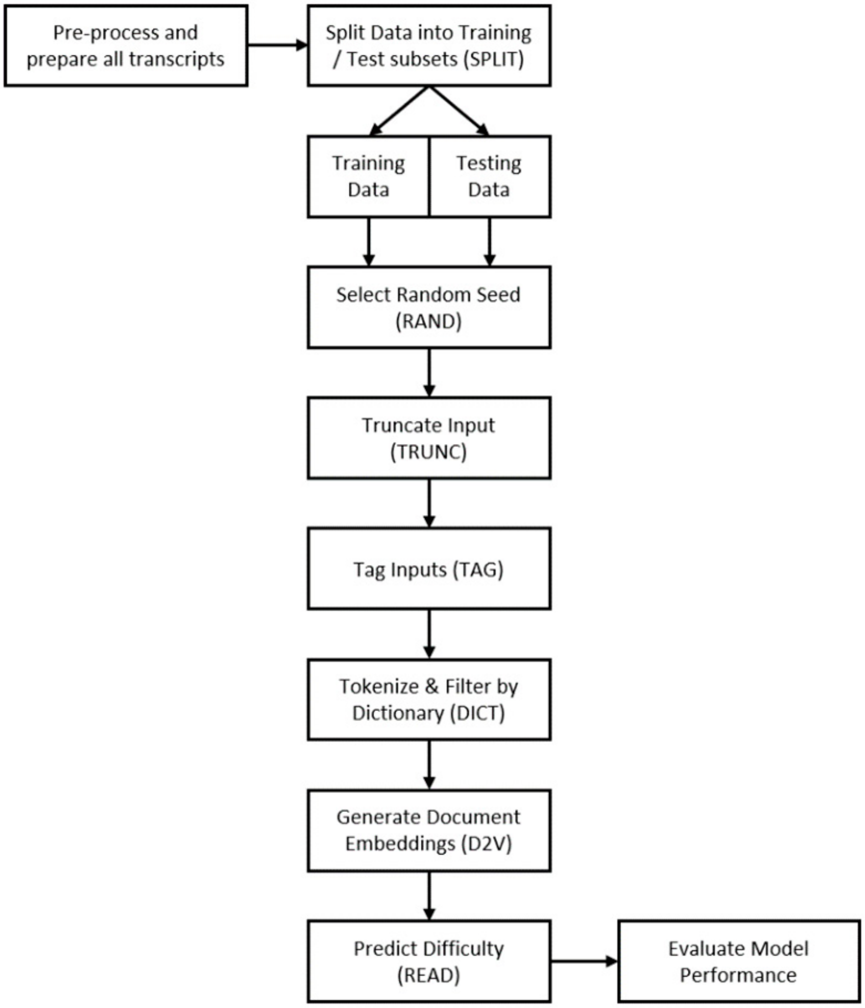
Institutional Review Board (IRB) approval to use this dataset for research was granted at Georgia State University where the study was exempted from further review. The IRB at Kansas State University also granted approval to use this dataset for research and exempted the study from further review.

All modeling, experimental design, and analyses were conducted using the Python programming language. The Pandas and SciPy modules were used extensively to prepare and transform data and produce descriptive statistical information.[23] The Scikit-Learn and Gensim packages were used extensively for text processing and modeling.[24] Statistical plots were generated using the Seaborn package.[25] For full code used to conduct analysis, see supplementary code file.

Overall Model Structure and Experimental Design

To experiment with and evaluate the utility of developing predictive models, the authors developed a multistage modeling process (see figure 1) that would preprocess all VRS transcripts into a form suited for text analysis and splitting the data into "training" and "test" subsets (SPLIT). Then all VRS transcripts were processed according to a variety of modeling parameters. For the

FIGURE 1
Overall Modeling Process


Pre-process and prepare all transcripts
Split Data into Training / Test subsets (SPLIT)
Training Data
Testing Data
Select Random Seed (RAND)
Truncate Input (TRUNC)
Tag Inputs (TAG)
Tokenize & Filter by Dictionary (DICT)
Generate Document Embeddings (D2V)
Predict Difficulty (READ)
Evaluate Model Performance

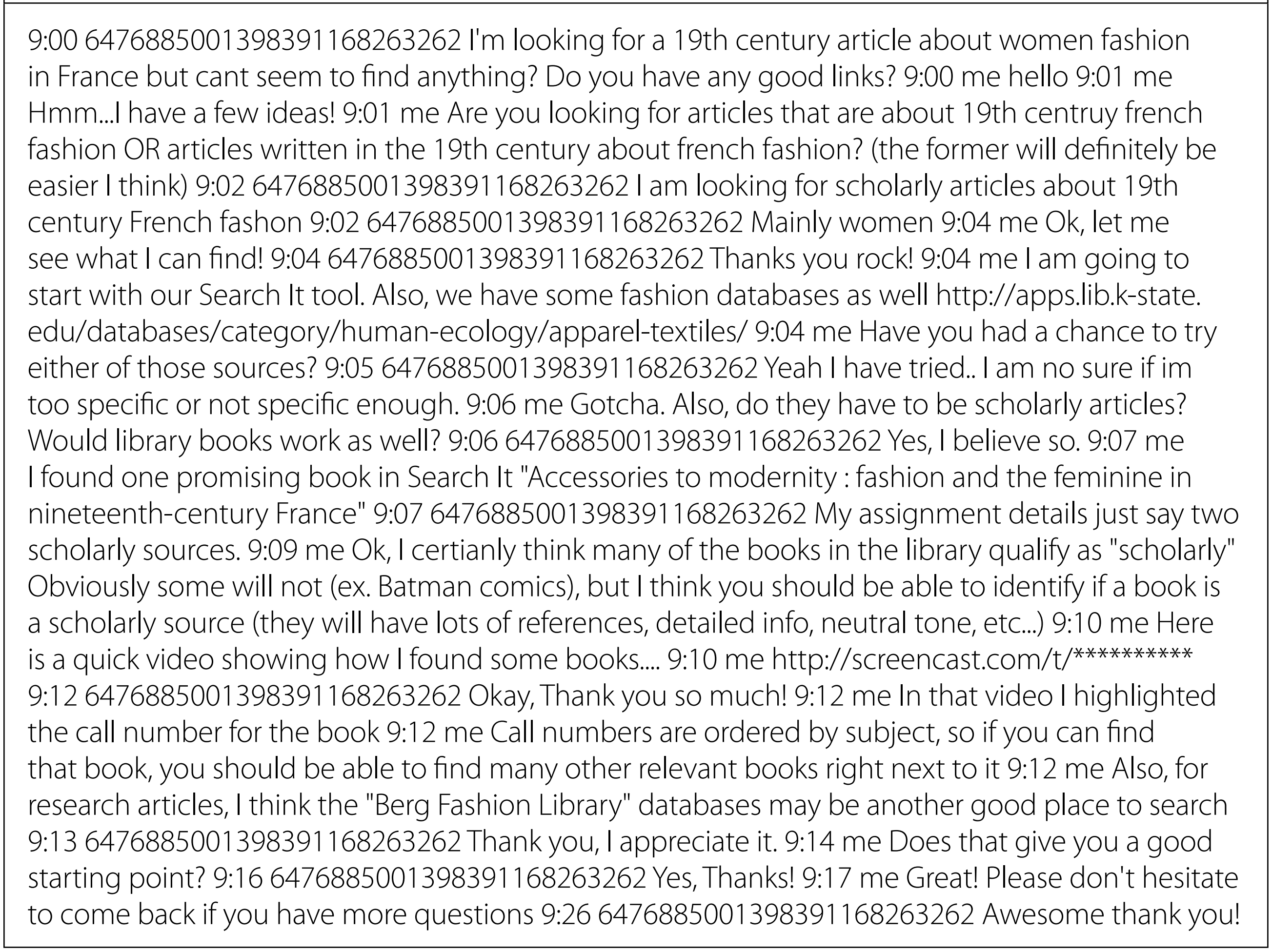
entire dataset, the entire modeling process was run one time for every possible combination of modeling parameters, resulting in predictive analytics data for 640 unique models.

In the overall modeling process and core experimental design, individual transcripts were truncated to just the few words of patron-supplied text (TRUNC), tagged with qualitative labels (TAG), modified to filter out infrequent terms (DICT), transformed into document-embeddings with fixed-length vector representations (D2V), and then processed through a very simple neural network classifier that would predict the relative *difficulty* of the VRS interaction using transformed READ Scale ratings (READ) as the formal representation *difficulty* of and final dependent variable in the modeling process.

Since some of the modeling processes and algorithms required randomly initialized states, a variable random parameter (RAND) was used to seed the random state for each individual run of the modeling process. Ultimately the final predictions for both the training-data and test-data for each individual model were evaluated using receiver operating characteristics' (ROC) area under the curve (AUC).[26] The AUC scores provided the core performance metric for evaluating the predictive power of individual model formulations.

Prior to modeling, the raw transcripts required extensive work to transform into a usable format. The raw, unformatted transcripts were exported from LibInsight and stored as plain text in the form shown in figure 2. Code was then written to automatically parse and organize each transcript into the form shown in figure 3.

FIGURE 2
Raw VRS Transcript



9:00 6476885001398391168263262 I'm looking for a 19th century article about women fashion in France but cant seem to find anything? Do you have any good links? 9:00 me hello 9:01 me Hmm...I have a few ideas! 9:01 me Are you looking for articles that are about 19th centruy french fashion OR articles written in the 19th century about french fashion? (the former will definitely be easier I think) 9:02 6476885001398391168263262 I am looking for scholarly articles about 19th century French fashon 9:02 6476885001398391168263262 Mainly women 9:04 me Ok, let me see what I can find! 9:04 6476885001398391168263262 Thanks you rock! 9:04 me I am going to start with our Search It tool. Also, we have some fashion databases as well http://apps.lib.k-state.edu/databases/category/human-ecology/apparel-textiles/ 9:04 me Have you had a chance to try either of those sources? 9:05 6476885001398391168263262 Yeah I have tried.. I am no sure if im too specific or not specific enough. 9:06 me Gotcha. Also, do they have to be scholarly articles? Would library books work as well? 9:06 6476885001398391168263262 Yes, I believe so. 9:07 me I found one promising book in Search It "Accessories to modernity : fashion and the feminine in nineteenth-century France" 9:07 6476885001398391168263262 My assignment details just say two scholarly sources. 9:09 me Ok, I certianly think many of the books in the library qualify as "scholarly" Obviously some will not (ex. Batman comics), but I think you should be able to identify if a book is a scholarly source (they will have lots of references, detailed info, neutral tone, etc...) 9:10 me Here is a quick video showing how I found some books.... 9:10 me http://screencast.com/t/********** 9:12 6476885001398391168263262 Okay, Thank you so much! 9:12 me In that video I highlighted the call number for the book 9:12 me Call numbers are ordered by subject, so if you can find that book, you should be able to find many other relevant books right next to it 9:12 me Also, for research articles, I think the "Berg Fashion Library" databases may be another good place to search 9:13 6476885001398391168263262 Thank you, I appreciate it. 9:14 me Does that give you a good starting point? 9:16 6476885001398391168263262 Yes, Thanks! 9:17 me Great! Please don't hesitate to come back if you have more questions 9:26 6476885001398391168263262 Awesome thank you!

FIGURE 3
Parsed VRS Transcript

```
['9:00', 'patron', "I'm looking for a 19th century article about women fashion in France but cant seem to find anything? Do you have any good links?"]
['9:00', 'staff', 'hello']
['9:01', 'staff', 'Hmm...I have a few ideas!']
['9:01', 'staff', 'Are you looking for articles that are about 19th centruy french fashion OR articles written in the 19th century about french fashion? (the former will definitely be easier I think)']
['9:02', 'patron', 'I am looking for scholarly articles about 19th century French fashon']
['9:02', 'patron', 'Mainly women']
['9:04', 'staff', 'Ok, let me see what I can find!']
['9:04', 'patron', 'Thanks you rock!']
['9:04', 'staff', 'I am going to start with our Search It tool. Also, we have some fashion databases as well http://apps.lib.k-state.edu/databases/category/human-ecology/apparel-textiles/']
['9:04', 'staff', 'Have you had a chance to try either of those sources?']
['9:05', 'patron', 'Yeah I have tried.. I am no sure if im too specific or not specific enough.']
['9:06', 'staff', 'Gotcha. Also, do they have to be scholarly articles? Would library books work as well?']
['9:06', 'patron', 'Yes, I believe so.']
['9:07', 'staff', 'I found one promising book in Search It "Accessories to modernity : fashion and the feminine in nineteenth-century France"']
['9:07', 'patron', 'My assignment details just say two scholarly sources.']
['9:09', 'staff', 'Ok, I certianly think many of the books in the library qualify as "scholarly" Obviously some will not (ex. Batman comics), but I think you should be able to identify if a book is a scholarly source (they will have lots of references, detailed info, neutral tone, etc...)']
['9:10', 'staff', 'Here is a quick video showing how I found some books....']
['9:10', 'staff', 'http://screencast.com/t/**********']
['9:12', 'patron', 'Okay, Thank you so much!']
['9:12', 'staff', 'In that video I highlighted the call number for the book']
['9:12', 'staff', 'Call numbers are ordered by subject, so if you can find that book, you should be able to find many other relevant books right next to it']
['9:12', 'staff', 'Also, for research articles, I think the "Berg Fashion Library" databases may be another good place to search']
['9:13', 'patron', 'Thank you, I appreciate it.']
['9:14', 'staff', 'Does that give you a good starting point?']
['9:16', 'patron', 'Yes, Thanks!']
['9:17', 'staff', "Great! Please don't hesitate to come back if you have more questions"]
['9:26', 'patron', 'Awesome thank you!']
```

FIGURE 4
Full Patron-supplied Text

I'm looking for a 19th century article about women fashion in France but cant seem to find anything? Do you have any good links? I am looking for scholarly articles about 19th century French fashon Mainly women Thanks you rock! Yeah I have tried.. I am no sure if im too specific or not specific enough. Yes, I believe so. My assignment details just say two scholarly sources. Okay, Thank you so much! Thank you, I appreciate it. Yes, Thanks! Awesome thank you!

Finally, all of the text provided by patrons was reassembled into a single text-string to represent the patron's contribution to the VRS interaction (see figure 4). To preserve the imperfect nature of text data in a chat environment, no attempt was made to correct typos, misspellings, or other linguistic errors of any kind.

While all VRS operators at KSUL are expected to redact personally identifying information from the text of these transcripts prior to submission to LibInsights, some human error is present in that process. To remedy this, part of the data-processing stage involved using regular expressions to identify instances of personal information (such as emails, phone numbers, "My name is ____"), redact targeted texts, and replace redactions with placeholders (such as "nameredacted," "emailredacted") to preserve the structure of the text while protecting individuals.

Following all data preparation procedures, some VRS transcripts remained that could not be incorporated into the research process. These samples were removed because they did not contain an original transcript, the transcript consisted entirely of redacted information, or the transcript was a duplicate of another entry. Once these samples were removed from the dataset, 14,604 individual transcripts remained for evaluation and modeling.

Once all usable transcripts were prepared and reduced down to just the patron-supplied text, the dataset was split into two subsets (SPLIT): "training" and "test." The SPLIT parameter was set at 2,000 (meaning that the most recent 2,000 VRS transcripts were placed into the test-data subset and the remaining 12,604 transcripts were placed into the training-data subset). The training data was used to define and fit the underlying components of various stages in the modeling process (TRUNC, DICT, D2V, and READ). Then, by evaluating the final predictive accuracy of each individual model using both the training data and test data, the authors evaluated both the internal and external validity of the modeling process.

Model: Parameters

In any NLP or text-driven task, researchers are faced with a wide variety of modeling choices. For instance, in the context of developing models for predicting VRS chat *difficulty*, one open question pertains to how READ Scale ratings should be grouped and labeled as "easier" or "harder." Since no model is perfect, segmenting READ Scale ratings at lower or higher thresholds will necessarily affect model performance—specifically, the tradeoff between precision and recall. An individual library service manager choosing to implement these types of models will need to set the threshold according to their preferences with respect to different measures of predictive accuracy. Ultimately, when deciding how to segment READ Scale ratings, as well as making choices for many other modeling parameters, identifying the "best" set of decisions is subjective and dependent on individual library service managers' strategic objectives.

To that end, attempting to find any single "best" model formulation was beyond the scope of the research project. Instead, the authors made multiple decisions for each of the primary modeling parameters and tested every possible combination of model formulation. This approach enabled the authors to evaluate and aggregate model performance metrics from a large sample of models that were independent of one another but still related through shared characteristics. A total of 640 unique individual models were created and evaluated. The model parameters used in the models are indicated in table 1.

- RAND—Random seed (RAND_0, … , RAND_19)

TABLE 1
Modeling Parameter Options

Parameter	Options	
RAND	0, 1, … , 19	
TRUNC	10	20
TAG	0	1
DICT	Full	Narrow
D2V	75	150
READ	1v2	2v3

At two later stages in the modeling process, specifically those pertaining to the D2V and READ parameter options, the models used neural-network processes to fit the model to the data. For these processes to function, the models required matrices containing randomly generated values. As each model was trained and fitted to the data, the values in the matrices were iteratively updated until the values converged at an optimal set of values.

However, because it was impossible to know if the values for any one model converged at truly optimal sets of values, as opposed to converging at a suboptimal local-minima, it was possible that an individual model could perform abnormally well or poorly due to random chance. To counter this, the authors incorporated the RAND parameter to manually change and control the random "seed," or initial value, for each model. Twenty unique values were used for the RAND parameter. Consequently, for every combination of the remaining modeling parameters, each model was refit and evaluated 20 times using a distinct random seed. This ensured that experimental results were robust to the effects of outliers.

- TRUNC—Truncating text inputs (TRUNC_10 and TRUNC_20)

For each sample, the input strings were truncated to either the first 10 or 20 space-delimited tokens. This functioned to provide the model with a minimal amount of text data to ensure model predictions were based strictly on the first small pieces of information provided by the patron at the outset of a VRS interaction. Drawing on the example in figure 4, the following examples show how the TRUNC parameter functioned to change the amount of text data that was input into the model for each sample:

  - TRUNC_10 input example: "I'm looking for a 19th-century article about women's fashion"
  - TRUNC_20 input example: "I'm looking for a 19th-century article about women's fashion in France but can't seem to find anything? Do you"

- TAG—Tagging documents (TAG_0 and TAG_1)

Following truncation, the input string for each sample was searched for the presence of specific patterns (such as "print in color") and labeled with corresponding tags (such as "tagPRINTING"). The authors used the tag-labels and patterns that were outlined in prior work.[27] Each individual input string could be labeled with zero, one, or multiple tags as appropriate. In some instances, the pattern represented literal strings of text. In other instances, the patterns were defined computationally using "regular expressions" as implemented in the Python programming language.[28]

Table 2 shows three examples of the labels for individual tags, the pattern-matching criteria that had to be present in the truncated text, and the conceptual definition for each tag. For a full listing of applied tags and associated text patterns, see appendix.

TABLE 2
TAG Examples

TAG	PATTERN	DEFINITION
tagPRINTING	print in color	Interactions relating to using library printers and printing services.
	cat cash	
tagQUIET	Quiet Zone	Interactions in which the user mentions excessive noise or inquires about quiet places in the library.
	floor to be quiet	
tagKNOWNITEMBOOK	a-z{12}\d{24}\s{01}\.a-z\d{1}	Interactions in which a user identifies a specific, individual book by name or Call Number.
	this book	

In model formulations using TAG_1, tags were used to supplement the training of document-embedding models and the generation of individual document-embeddings (D2V) for individual samples. For model formulation using TAG_0, the tags were not incorporated into the training of document-embedding models or in generating individual document-embeddings.

- DICT—Eliminating infrequent words (DICT_F and DICT_N)

Following tagging, the input string for each individual sample was tokenized into an array. In each input string, the individual tokens (that is to say, "words") were segmented by spaces and nonalphanumeric characters. All tokens were converted to lowercase letters in the process.

Once each input string was tokenized, all individual tokens remained in the input data for models using DICT_Full (DICT_F). For models using DICT_Narrow (DICT_N), the set of allowed tokens, or dictionary, was reduced by eliminating tokens that were too short or too infrequent across all of the input strings in the training data. To achieve this, the following filtering rules were implemented for models using the DICT_N parameter:

- Retain only alphanumeric tokens with a minimum length of 3 characters.
- Retain only tokens present in at least two transcripts in the training data.
- Retain only top 3,000 tokens as represented by average TF-IDF scores in the training data.

For the final filtering rule, TF-IDF ("term frequency—inverse document frequency") was used to count and then inversely weight the count-values against the proportion of input strings that contained any given individual token. This approach to assigning values to individual tokens is widely used for document classification tasks because it reduces the value of tokens that are extremely common, but unlikely to carry significant semantic value in isolation (such as "the"). The authors used the Scikit-Learn implementation of the TF-IDF method.[29]

The following examples illustrate how this process transformed each individual input string into a tokenized array as appropriate for both DICT_F and DICT_N models:

- TRUNC_10 input string:
  "I'm looking for a 19th century article about women fashion"
- DICT_F tokenized array:
  [i , m , looking , for , a , 19th , century , article , about , women , fashion]
- DICT_N tokenized array:
  [looking , 19th , century , article , about , women , fashion]

- D2V—Document Embeddings (D2V_75 and D2V_150)

Once the tokenized array for each individual transcript was set, a document-embedding model was trained using the training-data and the Gensim implementation of the Doc2Vec "PV-DBOW" algorithm.[30] The core outcome of a Doc2Vec algorithm and model is to transform tokenized arrays of input strings into fixed-length numeric representations of the data in numeric form. The algorithm achieves this by assigning a vector of values to each individual sample and then updating those values iteratively until samples that contain similar tokens will have similar vector representations.[31] At this stage in the modeling process, the variable modeling parameter was determining the length of the vectors that were used to represent the data. The authors chose to experiment with two different vector sizes: 75 and 150. For any given sample, the transformation of the tokenized array may look like the following example:

- □ DICT_N tokenized array:
  [looking , 19th , century , article , about , women , fashion]
- □ D2V_75 document embedding:
  $[\, 0.55_1 \,, -0.15_2 \,, 1.23_3 \,, \ldots \,, 1.77_{75} \,]$
- □ D2V_150 document embedding:
  $[\, 0.25_1 \,, -0.35_2 \,, -1.41_3 \,, \ldots \,, 1.57_{150} \,]$

An added perk of using the Gensim implementation of Doc2Vec is it allows for the individual document-embeddings to be trained alongside labels and tags. Consequently, as established in prior work, for model formulations using TAG_1, the document embeddings for individual samples that shared the same tags would be more closely aligned with each other than with other samples in the data.[32]

- READ—Dependent Variable (READ_1v2 and READ_2v3)

Finally, the dependent variable for the entire modeling process was defined using READ Scale ratings accompanying most, but not all, of the VRS transcripts in the dataset. Rather than trying to implement an ordinal statistical model, the authors decided to collapse the READ values into binary indicators using two different cutoffs. For READ_1v2, the original READ values were segmented between READ Scale ratings 1 and 2. For READ_2v3, the values were segmented between READ Scale ratings of 2 and 3. Subjectively, these splits transformed the original READ Scale values into two broad categories: "easier" (0) and "harder" (1) (see table 3).

TABLE 3
READ Scale Transformations

READ Scale	1	2	3	4	5	6
READ_1v2	0	1				
READ_2v3	0		1			

Using a simple neural network classifier as implemented in the Scikit-Learn package,[33] the document-embeddings for the training data were used to train and fit a binary classification model using the transformed READ values as the dependent variable. Finally, both the training data and test data were processed through the fully fitted classifier to generate predictions for each individual sample in both subsets. The final predictions for each sample were represented by probabilities ranging between 0.0 and 1.0.

Model—Evaluation

Following the full processing of samples in both the training data and test data, the overall performance of each model formulation was evaluated using the receiver operating characteristics (ROC) area under the curve (AUC) scores. When using a set of predicted probabilities, such as those generated by the described models, the ROC curve represents the change in the false-positive and true-positive classification rates as the decision threshold used for determining which probabilities are rounded down to zero (0) or up to one (1) is varied. When measuring the area under the ROC curve (AUC), AUC scores can range from 0.0 to 1.0 and in effect represent the overall discriminatory power of a model. For a binary classification task, AUC scores of 1.0 indicate that the model is able to perfectly discriminate between the two classes (that is to say, "easier" vs "harder"), for all possible decision thresholds. An AUC score of 0.5 would indicate that the model's capacity to discriminate between two classes is no better than randomly guessing.

By using AUC scores to evaluate each individual model's performance with respect to both the testing data and training data, the authors were able to inspect and evaluate the AUC scores in aggregate and as distributions. Additionally, simple right-tail t-tests were conducted to determine if the distribution of AUC scores demonstrated that the modeling techniques used were able to generate better-than-random predictions in a statistically significant way. The same evaluation of AUC scores was conducted on subsets of VRS transcripts with respect to *Question Type* labels that were recorded by VRS operators. Further, simple classification accuracy of each individual model was calculated with respect to TAG labels and relevant subsets of transcripts.

Results

The results of the analysis unambiguously demonstrated that the models are fully capable of providing library service managers with robust and reliably better-than-random predictions regarding the relative *difficulty* of inbound VRS chat sessions. In total, 640 individual models

TABLE 4
Mean AUC Scores by Data Subsets

	TRAINING DATA						**TESTING DATA**					
		AUC Scores						*AUC Scores*				
Data Subsets	*N (transcripts)*	*Min*	*Max*	*Mean*	*Std Dev*	*Significance*	*N (transcripts)*	*Min*	*Max*	*Mean*	*Std Dev*	*Significance*
All Samples	10,162	0.7207	0.8247	0.7700	0.0271	***	1,753	0.6220	0.7174	0.6631	0.0163	***
All untagged samples	8,330	0.7003	0.8186	0.7558	0.0298	***	1,413	0.5914	0.6957	0.6434	0.0202	***
All tagged samples	1,832	0.7746	0.8514	0.8110	0.0160	***	340	0.6567	0.7571	0.7116	0.0167	***
Question Type												
Building	709	0.5898	0.8391	0.7284	0.0478	***	70	0.4836	1.0000	0.7603	0.1743	***
Circulation	473	0.6054	0.8133	0.7063	0.0403	***	88	0.3460	0.7128	0.5305	0.0555	***
Copyright	8	0.2857	1.0000	0.8964	0.1509	***	2	0.0000	1.0000	0.0406	0.1977	
Directional	200	0.6526	0.9889	0.8145	0.0948	***	35	0.3750	1.0000	0.6324	0.1151	***
KAPI	6	0.0000	1.0000	0.7109	0.1656	***	0	—	—	—	—	—
KREx	12	0.1000	1.0000	0.7238	0.1996	***	2	—	—	—	—	—
Misc	917	0.6285	0.8060	0.7038	0.0367	***	172	0.5002	0.7894	0.6270	0.0533	***
NewPrairiePress	6	0.0000	1.0000	0.5989	0.2694	***	2	0.0000	1.0000	0.5313	0.4998	
Reference	6,743	0.6478	0.7883	0.7168	0.0339	***	1,167	0.5591	0.6778	0.6079	0.0216	***
ResearchConsultation	45	0.5427	0.9701	0.7447	0.0826	***	10	0.0000	1.0000	0.5360	0.2175	***
Reserves	110	0.5297	0.8105	0.6568	0.0542	***	10	0.0000	1.0000	0.5669	0.3155	***
Technical	907	0.6588	0.8581	0.7528	0.0392	***	179	0.4642	0.7701	0.6175	0.0708	***
Unknown	26	0.0400	1.0000	0.6126	0.2119	***	16	0.0000	1.0000	0.6274	0.2100	***

*** p < 0.001
** p < 0.01
* p < 0.05
" " Not statistically significant
"—" Insufficient sample size

with unique modeling parameters were evaluated. Table 4 shows a comprehensive summary of the AUC scores with respect to various subsets of the data.

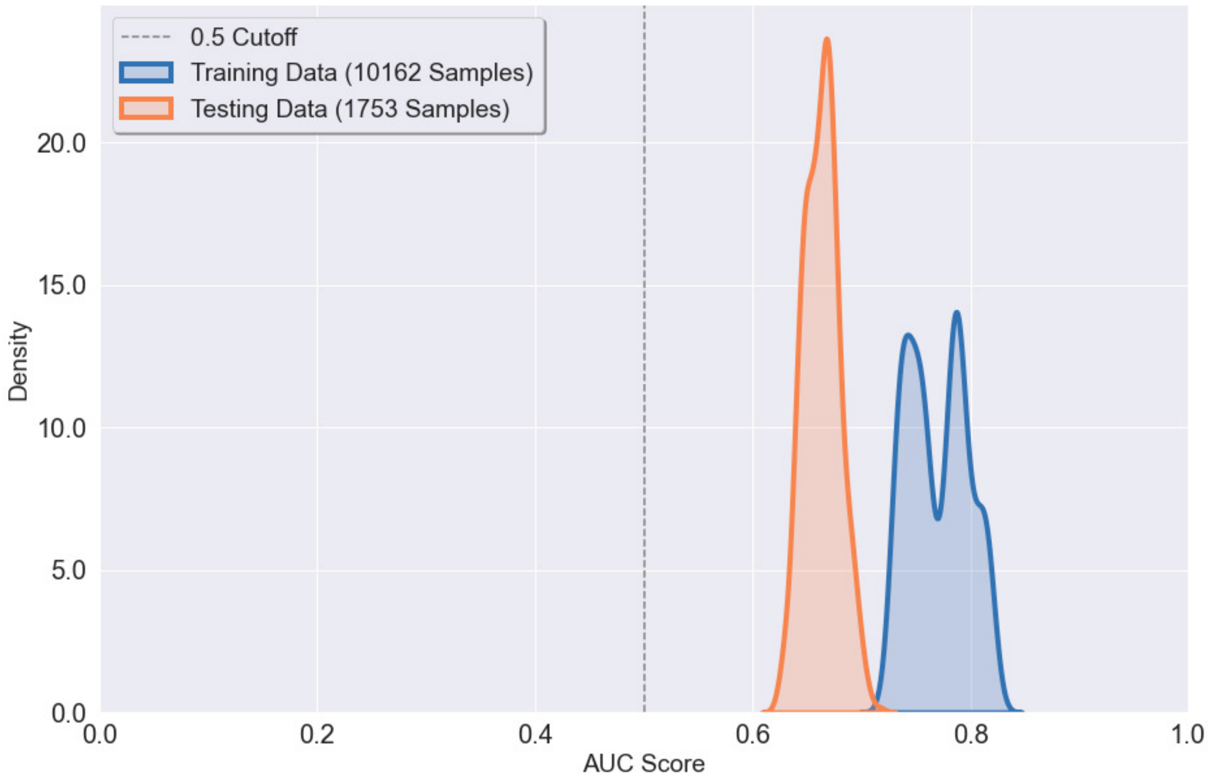
Since the final stage of the modeling process required fully valid samples to calculate AUC scores, only training and test data with fully labeled *READ* values were evaluated. Consequently, the number of VRS transcripts that received predictions and whose predictions were evaluated for final AUC scores from the training and testing data was reduced from 12,604 to 10,162 and from 2,000 to 1,753, respectively.

One aspect of the results that became immediately apparent was the discrepancy between the AUC scores for comparable training and testing subsets. For the overall model evaluations, the mean AUC score for the training data (in other words, internal validation) was approximately 0.7700, whereas the score for the testing data (that is, external validation) was 0.6631.

This pattern repeated itself in the results for all subsets of data. However, right-tailed t-tests revealed that the AUC scores for all models associated with the training data and nearly all models associated with the testing data were, on average, greater than 0.5 to a highly statistically significant degree ($p < 0.001$). Since the evaluation of the testing data was meant to simulate "new" VRS interactions, the results showed that the predictive models were reliably better than random at predicting the relative *difficulty* of individual VRS interactions.

In limited instances, the AUC scores for some testing subsets did not show statistically significant or better than random predictive accuracy. However, these subsets were also characterized by relatively small numbers of VRS samples (such as "Copyright," "KAPI," "KREx," and "NewPrairiePress"). Since the sample sizes associated with these subsets were so drastically small, AUC scores are not a reliable performance metric and, in some instances, were not available at all due to the homogeneity of the samples' *difficulty* ratings.

FIGURE 5
Distribution of AUC Scores for Full Training and Testing Datasets



Whereas the metrics in table 4 represent and characterize the AUC scores in aggregate, the AUC scores for all 640 models may be better understood as a distribution of modeling performance. For example, figure 5 highlights the distribution of AUC scores for all of the models with respect to the entirety of the training and testing subsets of the data.

Tagged and Untagged Subsets

As part of the modeling process, the authors were able to analyze AUC scores with respect to specific subsets of both the training and testing data. First, independent of whether the TAG modeling parameter was used during the training stage for any individual model, the predictions associated with VRS transcripts containing words, phrases, and patterns that matched any of the defined TAG labels as a distinct subset ("All tagged samples") were identified and evaluated. Additionally, all remaining samples were evaluated as a complementary subset ("All untagged samples").

The results show that, for both the training and testing data, there is a notable difference in the AUC scores between tagged and untagged VRS transcripts. For the training data, the AUC score for the tagged subset is greater than the untagged subset by approximately +0.0552. For the testing data, the same comparison shows a difference of +0.0681. For both of these comparisons, a two-sample t-test revealed that the difference in mean AUC scores was highly statistically significant ($p < 0.001$) and not a product of random chance.

In a very loosely defined sense, these results can be described as a 5 and 6 percent increase to the average predictive accuracy of the models when dealing with tagged samples in the training and testing data, respectively. These results demonstrated that VRS transcripts that are tagged are, almost by definition, easier to predict. This is due to the fact that the TAG labels, while subjectively defined, reflect common text patterns that librarians can easily identify and characterize. It should be taken as an encouraging sign that the tagged samples contained patterns that the modeling processes were able to detect and leverage to ultimately produce higher-quality predictions as reflected in the increased AUC scores, even in instances when the modeling parameters did not explicitly include TAG labels as part of the model fitting and training regime.

Question Types

In KSUL's LibInsight dataset, every recorded patron interaction, VRS transcripts included, has a *Question Type* label. These subjective nominal labels enabled the authors to further investigate model performance with respect to a variety of subsets of the data. Although some of the *Question Type* labels are idiosyncratic to KSUL (such as "KAPI," "KREx," and "NewPrairiePress"), other labels represent categories of questions that are common across library settings (such as "Directional" and "Reference").

One issue that stood out in the results (see table 4) was that *Question Type* labels "KAPI" and "KREx" were not associated with enough samples in the testing data to calculate AUC scores. Furthermore, the "Copyright" and "NewPrairiePress" labels were associated with so few samples in the testing data that the mean AUC scores, while positive, were not remotely statistically significant ($p > 0.10$). The mean AUC scores for all other subsets associated with the remaining *Question Type* labels were statistically significant ($p < 0.001$) and showed that the modeling processes could reliably predict the relative *difficulty* of VRS transcripts across multiple categories.

Another interesting result that manifested in the data was the discrepancy between the respective *Question Type* subsets in the training and testing data. As shown in figure 6, the

FIGURE 6
Distribution of AUC Scores by Question Type (Training Data)

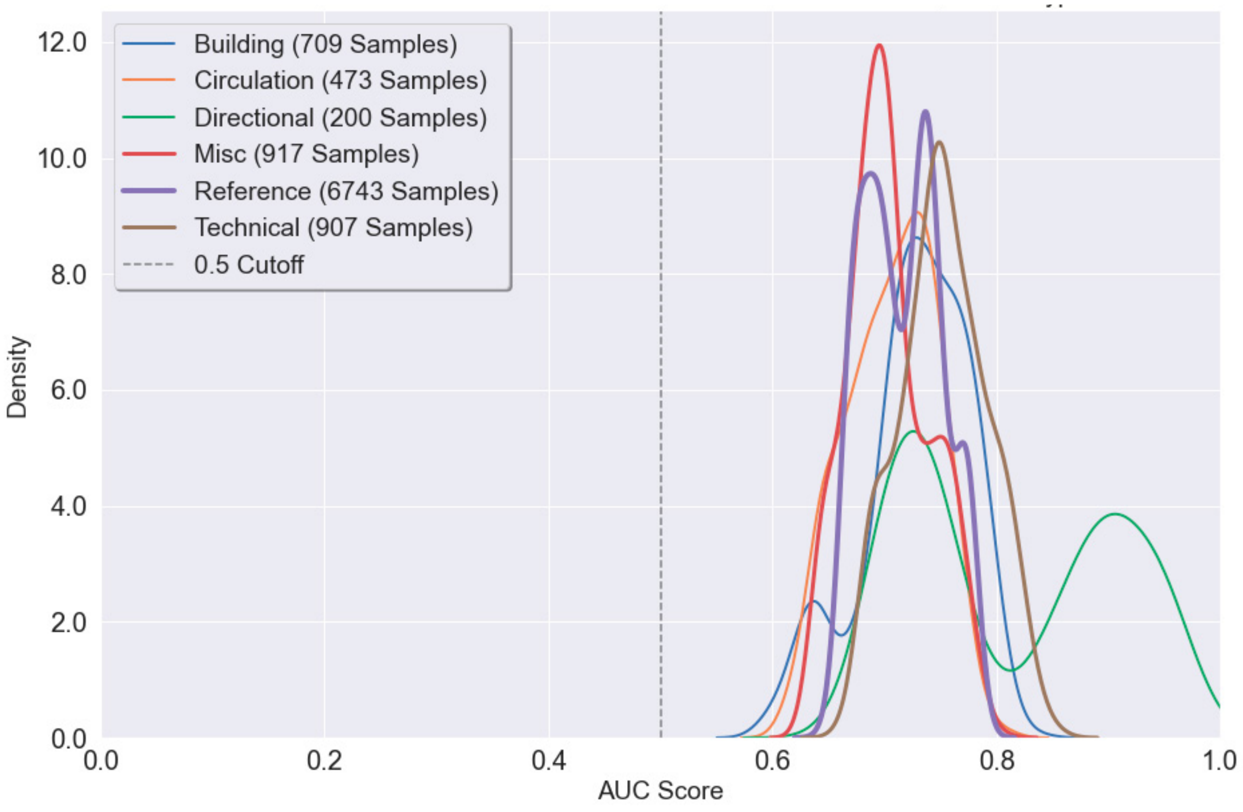
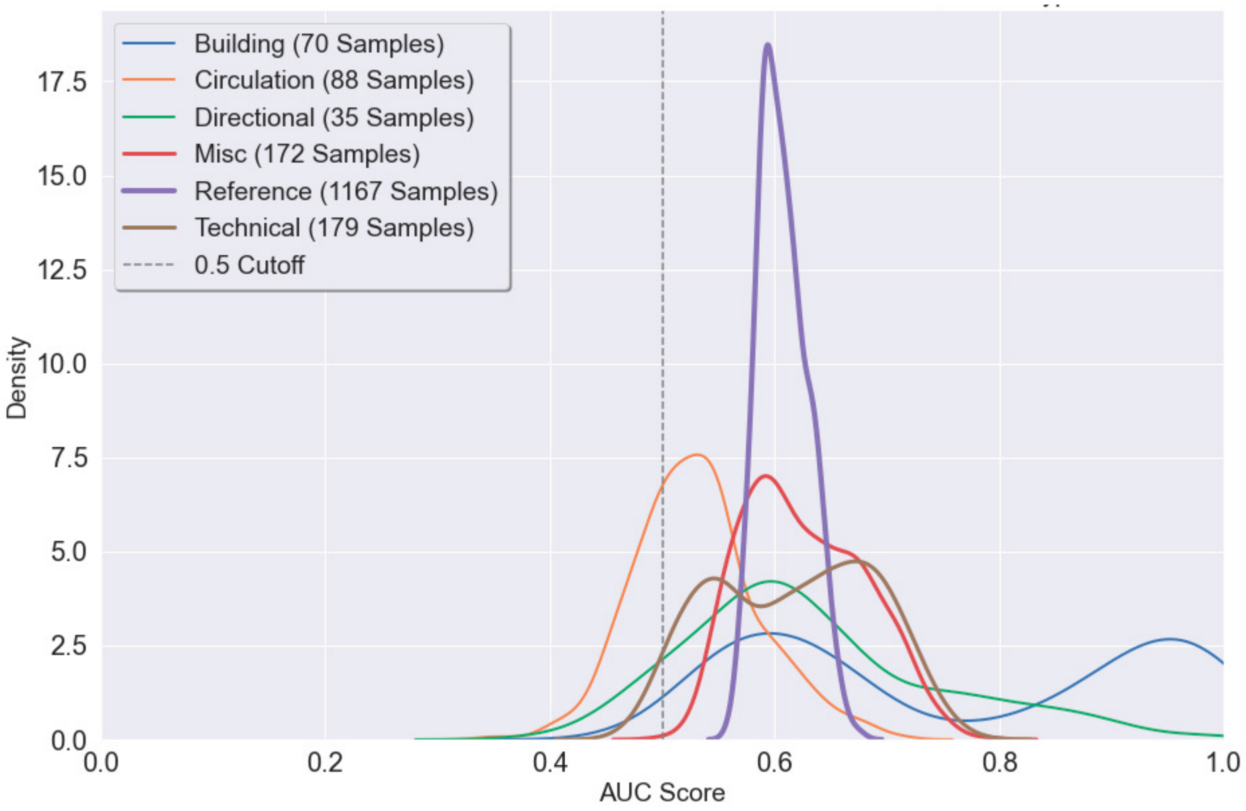


FIGURE 7
Distribution of AUC Scores by Question Type (Testing Data)



distributions of the mean AUC scores in the training data for the subsets associated with the six most prevalent *Question Type* labels ("Building," "Circulation," "Directional," "Misc," "Reference," and "Technical") were extremely similar. By contrast, figure 7 shows the same distributions for the testing data. In the latter figure, the distributions for each of the *Question Type* labels were markedly different, indicating that the modeling processes do not perform consistently across all categories when evaluated against external or new data. With respect to the testing data, while the "Reference" questions maintained a relatively narrow and dense distribution range of AUC scores, the results associated with the other *Question Type* labels were much more varied and dispersed.

TAG Parameter Labels

While the authors were able to identify granular subsets of samples that matched patterns among individual TAG labels (see appendix) in both the training and testing data, it was not possible to evaluate these subsets using AUC scores. For many of these subsets, there were either too few samples or all available samples' READ labels were identical and thus could not be evaluated using ROC AUC as a performance metric. However, it was possible to calculate simple classification accuracy scores for nearly all subsets in both the training and testing data (see table 5).

TABLE 5
Mean Classification Accuracy Scores by TAG Labels

	TRAINING DATA			**TESTING DATA**	
TAG Labels	**N (transcripts)**	**Average Accuracy Score**		**N (transcripts)**	**Average Accuracy Score**
tagARTICLES	786	74.88%		141	69.21%
tagCURRICULUM	7	74.53%		1	52.97%
tagEVIDENCEBASED	20	77.62%		1	50.31%
tagHOURS	253	83.55%		62	72.74%
tagJUVENILE	40	81.20%		10	84.10%
tagKNOWNITEMARTICLE	258	73.90%		78	69.26%
tagKNOWNITEMBOOK	312	76.96%		57	75.92%
tagLIBHALE	521	78.79%		109	75.68%
tagLIBLOCATION	323	84.93%		38	73.22%
tagLIBMATHPHYS	6	76.38%		0	—
tagLIBSTACKS	50	79.94%		5	86.09%
tagLIBVETMED	8	65.25%		2	84.92%
tagLIBWEIGEL	15	76.33%		5	61.27%
tagPRINTING	119	82.38%		12	69.49%
tagQUIET	125	89.08%		15	68.56%
tagREFERENCE	381	74.92%		49	68.66%
tagSCANNER	49	82.28%		8	77.59%
tagTEXTBOOKS	90	77.16%		19	83.93%
tagURL	206	77.74%		52	72.83%

Similar to the aforementioned results, the predictive accuracy of the modeling processes was greater for the training data than for the testing data. However, even when predicting the relative *difficulty* of the VRS transcripts in the testing data, the predictive accuracy associated with many of the individual TAG labels indicated that the modeling processes generated robust predictions. For example, for the two most prevalent TAG labels in the testing dataset, "tagARTICLES" and "tagLIBHALE," the average classification accuracy score across all modeling processes was 69.21 and 75.68 percent, respectively.

Further, it is worth noting that the results also demonstrated that the modeling processes were able to, on average, produce equally robust predictions for VRS inquiries representing notably different types of library patrons' inquiries. Comparisons between the definition and scope of any two TAG labels reveals distinctions in types of patron inquiries characterized and detected in individual VRS transcripts. Despite these distinctions, the average classification accuracy across many of the TAG subsets are comparable.

In just one example drawn from the testing results, the average accuracy score associated with "tagARTICLES" and "tagKNOWNITEMARTICLE" subsets of the data were nearly the same (accuracy ~69%). The TAG label for the former was intended to identify VRS transcripts in which the patron was expressing an open-ended inquiry regarding how to find and access journal articles, while the latter represented transcripts in which the patron expressed the need for support in accessing a specific article that was already known to them.

Limitations and Future Directions

The scope of this research was limited to the exploration and evaluation of predictive modeling processes. This research was predicated on the assumption that this modeling approach could, hypothetically, be deployed by individual libraries seeking to systematically triage incoming VRS inquiries to appropriate library staff operators. Since this modeling approach has been shown to be distinctly better than random guessing, it is clear that the approach outlined here can provide developers and librarians with a useful roadmap for implementing and evaluating their own models.

However, it may be worth noting that the implementation of these models may be extremely challenging. From a technical standpoint, at the time of writing, none of the prominent VRS chat services offered by third-party software developers allow librarians to directly integrate decision or triage models into the functional aspects of their platforms. Consequently, to test any of the proposed modeling processes, librarians and developers will face either distinct engineering or business challenges—or both—just to implement a model in a live situation. For most libraries, this will likely represent a prohibitive barrier to further testing, experimentation, and, ultimately, improvements to library services.

Independent of the technical challenges associated with implementing a predictive model in the context of VRS services, many aspects of this research need to be investigated and modified locally if other libraries intend to develop, test, and deploy their own models. In this research, the authors experimented with a wide variety of modeling parameters, but libraries wishing to use this technology will need to seriously consider how to select and implement particular modeling parameters.

For example, the TAG parameter in these modeling processes was characterized by idiosyncratic and subjective labels that are largely only relevant to the characteristics of VRS transcripts at KSUL. Further, selecting appropriate TRUNC, DICT, D2V, and READ modeling parameters is highly contingent upon the characteristics of local data and library managers' service objectives. For many of these decisions, there are no generalizable "right" or "wrong" answers, just modeling decisions that reflect the local needs of individual libraries.

Making these decisions may present some librarians with a steep burden with respect to the combination of technical, analytic, and strategic planning skills necessary to evaluate modeling parameters and performance. This research design centered around testing a myriad of models using many combinations of modeling parameters. Consequently, to run the analyses in a reasonable period of time, a variety of decisions were consciously made that resulted in faster processing, fitting, and evaluation of models at the expense of the overall predictive power of each modeling process.

The options associated with nearly every step of the modeling process were, in effect, arbitrarily chosen. This provided useful experimental information and comparisons, but not necessarily the best possible individual model. Additionally, when training the document-embedding model within the overall modeling process (D2V), the fitting-step was limited to 100 iterations. The same limit was also imposed on the number of training iterations for the neural network classifier at the final stage of the modeling process. The consequences of these decisions resulted in the slowest model processing time for any individual model being approximately 130 seconds. Further, the maximum AUC scores across all models with respect to the training and testing subsets of the data was 0.8247 and 0.7174, respectively. While these models are significantly better than random (AUC = 0.5), these scores represent the extremes, and it can be assumed that models with significantly better and more robust AUC scores can be developed with additional model tuning.

To improve upon the model designs that were tested, future development will need to be focused on fine tuning many individual modeling parameters that are embedded through the modeling process. The stages of the process where models require fitting (such as document-embedding and classifier models) will require significantly more training iterations for the models to converge on optimal formulations. Further, all the processes described and evaluated in this research are highly sensitive to the size of the underlying dataset. With increasingly large numbers of VRS transcripts and fully labeled samples, future research and development should be able to greatly advance the predictive power of these types of models.

The last significant limitation that must be noted is that no model, no matter how finely tuned, will be perfect. Even if a library service manager can develop and implement a robust predictive model for the purposes of automatically triaging incoming VRS transcripts, there will always be a certain degree of error. As such, VRS operators will still need to be able to ask open-ended questions, conduct reference interviews, and refer VRS users to appropriate VRS operators when appropriate.

Conclusion

Based upon the results manifest from this research project, it is clear that the application of machine learning, NLP, and advanced modeling techniques in the context of academic library services represents a rich and unique avenue for librarians to improve and build upon existing services. If implemented, these models represent a robust method for saving costs by automatically and instantly routing virtual patron inquiries to appropriate library employees. This can directly translate to a more efficient use of librarians' time and labor. Additionally, this represents an opportunity to ensure that patrons experience a higher quality of service by being connected directly to library employees who are best suited to their needs. Although the research conducted in this study was focused on the relative *difficulty* of patron inquiries, these models can be easily retooled and generalized to triage incoming virtual reference inquiries using any other categorical variables for which there is sufficient data. While there is much more research and development that can and should be done in this area, the use of these advanced techniques represents a golden opportunity for libraries to improve upon and transform their services.

APPENDIX

TAG	PATTERN	DEFINITION
tags / labels attributed to individual samples	*substring or REGEX patterns to be detected in raw user-supplied texts*	*Subjective definition of the broad concepts associated with core-classes and sub-classes in relation to VRS interactions*
tagPRINTING	color print	Interactions related to using library printers and printing services.
	colored print	
	print in color	
	print something in color	
	\Win color\W	
	cat cash	
	printer	
	(?<!3D\s)\bprinting	
	double.{1}sided	
	catcash	
	cat cash	
	add money	
tagSCANNER	scanner	Interactions related to using library scanners.
	\Wscan\W	
tagHOURS	open 24/7	Interactions in which users inquire about library building and service hours.
	what time	
	the hours	
	opens{0,1}\W	
	will be open	
	summer hours	
	library hours	
tagLIBMATHPHYS	re.escape('Math/Physics Library')	Interactions in which the Math & Physics Library is explicitly identified.
	re.escape('math and physics library')	
	re.escape('Math Physic library')	
	re.escape('math/physics library')	
	re.escape('maths/phys library')	
	re.escape('math & phys library')	
	re.escape('math phys library')	
tagLIBWEIGEL	weigel	Interactions in which the Weigel Architecture Library is explicitly identified.
	wiegel	
tagLIBVETMED	vet med	Interactions in which the Veterinary Medicine Library is explicitly mentioned.
	vetmed	
tagLIBHALE	Hale Library	Interactions in which Hale Library is explicitly identified.
	(?<!help\s)hale	
tagLIBSTACKS	Library Stacks	Interactions in which the users explicitly identify the "stacks."
	the stacks	
	in Stacks	

TAG	PATTERN	DEFINITION
tags / labels attributed to individual samples	*substring or REGEX patterns to be detected in raw user-supplied texts*	*Subjective definition of the broad concepts associated with core-classes and sub-classes in relation to VRS interactions*
tagTEXTBOOKS	the reserve	Interactions in which textbooks and course reserve materials and services are mentioned.
	on reserve	
	course reserve	
	reserve textbook	
	have a specific textbook	
	have the textbook	
	have textbook	
	this textbook	
	this text book	
tagQUIET	quite loud	Interactions in which the user mentions excessive noise or inquires about quiet places in the library.
	super loud	
	really loud	
	very loud	
	stop talking	
	talking on	
	music loud	
	loud	
	talking very	
	talking extremely	
	talking loud	
	quiet floor	
	Quiet Zone	
	quiet floors	
	floor to be quiet	
	whisper quietly	
	be quiet	
	floor to be quiet	

TAG	PATTERN	DEFINITION
tags / labels attributed to individual samples	*substring or REGEX patterns to be detected in raw user-supplied texts*	*Subjective definition of the broad concepts associated with core-classes and sub-classes in relation to VRS interactions*
tagLIBLOCATION	first floor	Interactions in which the user mentions specific locations within the library (that is to say, Hale Library).
	1st floor	
	second floor	
	2nd floor	
	third floor	
	3rd floor	
	fourth floor	
	4t floor	
	fifth floor	
	5th floor	
	hemisphere room	
	Harry Potter room	
	the hemi	
tagKNOWNITEMBOOK	a-z{12}\d{24}\s{01}\.a-z\d{1}	Interactions in which a user identifies a specific, individual book.
	this book	
tagARTICLES	peer.{,1}review	Interactions in which users ask about accessing, finding, or discovering journal articles in general.
	journal article	
	scholarly article	
	scholarly journal	
	peer reviewed	
	re.escape('peer-reviewed')	
	peerreviewed	
	scholarly	
	articles	
tagEVIDENCEBASED	evidence.based	Interactions in which users explicitly ask about evidence-based biomedical and health sciences research.
	kinesiology	
tagJUVENILE	juv lit section	Interactions in which users ask about the Juvenile Literature collection or inquire about the availability of children's literature more broadly.
	Juvenile Literature	
	re.escape(juv. lit)	
	children{01}s collection	
	children{01}s lit	
	children{01}s stor	
	re.escape(childrens boooks)	
	(?<!Germany on English) children{01}s book	
	re.escape(childrens picture)	
	picture book	

TAG	PATTERN	DEFINITION
tags / labels attributed to individual samples	*substring or REGEX patterns to be detected in raw user-supplied texts*	*Subjective definition of the broad concepts associated with core-classes and sub-classes in relation to VRS interactions*
tagCURRICULUM	curriculum materials	Interactions in which users ask about the Curriculum Materials Center or K–12 education materials more broadly.
	curriculum books	
tagKNOWNITEMARTICLE	doi\W\s{1}\S+	Interactions in which a user identifies a specific, individual journal article.
	doi:{01}\s{01}\d\S+	
	this article	
	this\s\w+\sarticle	
	this paper	
	doi\.\S+	
	doi:{01}\s{01}\d\S+	
	doi\.org\S+	
tagREFERENCE	articles{01}\sabout	Interactions in which users ask broadly for reference/research support and guidance.
	books{01}\sabout	
	subject	
	topic	
	a paper on	
	help me find an{01}	
tagURL	re.escape(amazon.com)	Interactions in which a user shares a URL to any website.
	re.escape(newfirstsearch)	
	re.escape(galegroup)	
	re.escape(ingentaconnect.com)	
	re.escape(proquest.com)	
	re.escape(ncbi.nlm.nih.gov)	
	re.escape(sciencedirect.com)	
	re.escape(springer.com)	
	re.escape(tandfonline.com)	
	re.escape(webofknowledge)	
	re.escape(wiley.com)	
	re.escape(books.google)	
	re.escape(google.com)	
	re.escape(apps.lib.k-state.edu/databases)	
	re.escape(er.lib.ksu.edu)	
	re.escape(er.lib.k-state.edu)	
	re.escape(getit.lib.ksu.edu)	
	re.escape(getit.lib.k-state.edu)	
	re.escape(guides.lib.ksu.edu)	

TAG	PATTERN	DEFINITION
tags / labels attributed to individual samples	*substring or REGEX patterns to be detected in raw user-supplied texts*	*Subjective definition of the broad concepts associated with core-classes and sub-classes in relation to VRS interactions*
tagURL	re.escape(guides.lib.k-state.edu)	Interactions in which a user shares a URL to any website.
	re.escape(catalog.lib.ksu.edu)	
	re.escape(catalog2.lib.ksu.edu)	
	re.escape(catalog.lib.k-state.edu)	
	re.escape(catalog2.lib.k-state.edu)	
	re.escape(primo.hosted.exlibrisgroup.com)	
	re.escape(na02.alma.exlibrisgroup)	
	re.escape(searchit.lib.ksu.edu)	
	re.escape(searchit.lib.k-state.edu)	
	re.escape(lib.k-state.edu)	
	re.escape(lib.k-state.edu)	
	re.escape(doi.org)	
	re.escape(http)	
	re.escape(www.)	

Notes

1. Jeremy Walker and Jason Coleman, “Analyzing Virtual Reference Transcripts with Machine Learning,” in *Library Technology Conference* (2019), https://digitalcommons.macalester.edu/libtech_conf/2019/sessions/27; Jeremy Walker, “Measuring the Impact That Domain Knowledge in the Form of Simple Ontology Structures Has on Predictive Modelling Processes in the Context of Academic Library Virtual Reference Services” (M.S., Evanston, IL, Northwestern University, 2019), https://doi.org/10.21985/N20V2W.

2. David Ward, “Using Virtual Reference Transcripts for Staff Training,” *Reference Services Review* 31, no. 1 (March 1, 2003): 46–56, https://doi.org/10.1108/00907320310460915; Amy Burger, Jung-ran Park, and Guisu Li, “Application of Reference Guidelines for Assessing the Quality of the Internet Public Library’s Virtual Reference Services,” *Internet Reference Services Quarterly* 15, no. 4 (October 2010): 209–26, https://doi.org/10.1080/10875301.2010.526479; Greta Valentine and Brian D. Moss, “Assessing Reference Service Quality: A Chat Transcript Analysis” (March 28, 2017), https://kuscholarworks.ku.edu/handle/1808/25179; Deborah L. Meert and Lisa M. Given, “Measuring Quality in Chat Reference Consortia: A Comparative Analysis of Responses to Users’ Queries,” *College & Research Libraries* 70, no. 1 (January 2009): 71–84, https://doi.org/10.5860/0700071; Marie L. Radford, “Encountering Virtual Users: A Qualitative Investigation of Interpersonal Communication in Chat Reference,” *Journal of the American Society for Information Science and Technology* 57, no. 8 (2006): 1046–59, https://doi.org/10.1002/asi.20374; Koshik Irene and Okazawa Hiromi, “A Conversation Analytic Study of Actual and Potential Problems in Communication in Library Chat Reference Interactions,” *Journal of the American Society for Information Science and Technology* 63, no. 10 (September 10, 2012): 2006–19, https://doi.org/10.1002/asi.22677; Jennifer Waugh, “Formality in Chat Reference: Perceptions of 17- to 25-Year-Old University Students,” *Evidence Based Library and Information Practice* 8, no. 1 (March 1, 2013): 19–34; JoAnn Jacoby et al., “The Value of Chat Reference Services: A Pilot Study,” *portal: Libraries and the Academy* 16, no. 1 (February 18, 2016): 109–29, https://doi.org/10.1353/pla.2016.0013.

3. Yasmin Morais and Sara Sampson, “A Content Analysis of Chat Transcripts in the Georgetown Law Library,” *Legal Reference Services Quarterly* 29, no. 3 (July 1, 2010): 165–78, https://doi.org/10.1080/02703191003751289; Amanda Clay Powers, Julie Shedd, and Clay Hill, “The Role of Virtual Reference in Library Web Site Design: A Qualitative Source for Usage Data,” *Journal of Web Librarianship* 5, no. 2 (April 1, 2011): 96–113, https://doi.org/10.1080/19322909.2011.573279.

4. Miriam L. Matteson, Jennifer Salamon, and Lindy Brewster, "A Systematic Review of Research on Live Chat Service," *Reference & User Services Quarterly* 51, no. 2 (December 5, 2011): 172–89, https://doi.org/10.5860/rusq.51n2.172.

5. Marianne Stowell Bracke et al., "Finding Information in a New Landscape: Developing New Service and Staffing Models for Mediated Information Services," *College & Research Libraries* 68, no. 3 (2007), https://doi.org/10.5860/crl.68.3.248; Kate Fuller and Nancy H. Dryden, "Chat Reference Analysis to Determine Accuracy and Staffing Needs at One Academic Library," *Internet Reference Services Quarterly* 20, no. 3/4 (July 2015): 163–81, https://doi.org/10.1080/10875301.2015.1106999; Krisellen Maloney and Jan H. Kemp, "Changes in Reference Question Complexity Following the Implementation of a Proactive Chat System: Implications for Practice," *College & Research Libraries* 76, no. 7 (2015): 959–74, https://doi-org.turing.library.northwestern.edu/10.5860/crl.76.7.959; Valery King and Sara Christensen-Lee, "Full-Time Reference with Part-Time Librarians," *Reference & User Services Quarterly* 54, no. 1 (September 25, 2014): 34–43, https://doi.org/10.5860/rusq.54n1.34.

6. Patricia Bravender, Colleen Lyon, and Anthony Molaro, "Should Chat Reference Be Staffed by Librarians? An Assessment of Chat Reference at an Academic Library Using Libstats," *Internet Reference Services Quarterly* 16, no. 3 (July 1, 2011): 111–27, https://doi.org/10.1080/10875301.2011.595255.

7. Tara Radniecki and Mitch Winterman, "Leveraging Student Expertise for Niche Services," *Reference Services Review* 48, no. 2 (January 1, 2020): 287–306, https://doi.org/10.1108/RSR-11-2019-0083.

8. Vera J. Lux and Linda Rich, "Can Student Assistants Effectively Provide Chat Reference Services? Student Transcripts vs. Librarian Transcripts," *Internet Reference Services Quarterly* 21, no. 3/4 (July 2016): 115–39, https://doi.org/10.1080/10875301.2016.1248585; Kelsey Keyes and Ellie Dworak, "Staffing Chat Reference with Undergraduate Student Assistants at an Academic Library: A Standards-Based Assessment," *Journal of Academic Librarianship* 43, no. 6 (November 1, 2017): 469–78, https://doi.org/10.1016/j.acalib.2017.09.001.

9. Craig Boman et al., *Artificial Intelligence and Machine Learning in Libraries*, ed. Jason Griffey, vol. 55, Library Technology Reports 1 (Chicago, IL: ALA TechSource, 2019); University of Oklahoma Libraries, "Projects in Artificial Intelligence Registry (PAIR): A Registry for AI Projects in Higher Ed," Projects in Artificial Intelligence Registry (PAIR), 2019, https://pair.libraries.ou.edu/.

10. University of Oklahoma Libraries, "ANTswers: University of California, Irvine Libraries Chatbot," Projects in Artificial Intelligence Registry (PAIR), January 18, 2019, https://pair.libraries.ou.edu/content/antswers-university-california-irvine-libraries-chatbot; University of Oklahoma Libraries, "Library Website Chatbot," Projects in Artificial Intelligence Registry (PAIR), January 18, 2019, https://pair.libraries.ou.edu/content/library-website-chatbot.

11. University of Oklahoma Libraries, "ANTswers: University of California, Irvine Libraries Chatbot."

12. Walker and Coleman, "Analyzing Virtual Reference Transcripts with Machine Learning"; Walker, "Measuring the Impact That Domain Knowledge in the Form of Simple Ontology Structures Has on Predictive Modelling Processes in the Context of Academic Library Virtual Reference Services."

13. Ellie Kohler, "What Do Your Library Chats Say? How to Analyze Webchat Transcripts for Sentiment and Topic Extraction," in *Brick & Click Libraries Conference Proceedings* (17th Annual Brick & Click Libraries Conference, Maryville, Missouri: Northwest Missouri State University, 2017).

14. Kohler, "What Do Your Library Chats Say?"; Bella K. Gerlich and G. Lynn Berard, "Introducing the READ Scale: Qualitative Statistics for Academic Reference Services," *Georgia Library Quarterly* 43, no. 4 (2007).

15. Kohler, "What Do Your Library Chats Say?"

16. Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang, *Fundamentals of Predictive Text Mining* (New York, NY: Springer, 2015).

17. Quoc Le and Tomas Mikolov, "Distributed Representations of Sentences and Documents," in *International Conference on Machine Learning* (2014), 1188–96, http://proceedings.mlr.press/v32/le14.html.

18. Paula Lauren et al., "Discriminant Document Embeddings with an Extreme Learning Machine for Classifying Clinical Narratives," *Neurocomputing*, Hierarchical Extreme Learning Machines, 277 (February 14, 2018): 129–38, https://doi.org/10.1016/j.neucom.2017.01.117; M. Oubounyt et al., "Deep Learning Models Based on Distributed Feature Representations for Alternative Splicing Prediction," *IEEE Access* 6 (2018): 58826–34, https://doi.org/10.1109/ACCESS.2018.2874208; Roberta A. Sinoara et al., "Knowledge-Enhanced Document Embeddings for Text Classification," *Knowledge-Based Systems* 163 (January 1, 2019): 955–71, https://doi.org/10.1016/j.knosys.2018.10.026.

19. Jey Han Lau and Timothy Baldwin, "An Empirical Evaluation of Doc2vec with Practical Insights into Document Embedding Generation," in *Proceedings of the 1st Workshop on Representation Learning for NLP* (Berlin, Germany: Association for Computational Linguistics, 2016), 78–86, https://doi.org/10.18653/v1/W16-1609.

20. Natalya F. Noy and Deborah L McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Technical Report (Stanford Knowledge Systems Laboratory: Stanford University, March 2001), http://

www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html; Faisal Alshargi et al., "Concept2vec: Metrics for Evaluating Quality of Embeddings for Ontological Concepts," *ArXiv:1803.04488 [Cs]*, March 12, 2018, http://arxiv.org/abs/1803.04488.

21. Walker, "Measuring the Impact That Domain Knowledge in the Form of Simple Ontology Structures Has on Predictive Modelling Processes in the Context of Academic Library Virtual Reference Services."

22. Walker and Coleman, "Analyzing Virtual Reference Transcripts with Machine Learning."

23. Pauli Virtanen et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods* 17 (2020): 261–72, https://doi.org/10.1038/s41592-019-0686-2; Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, eds. Stéfan van der Walt and Jarrod Millman (2010), 56–61, https://doi.org/10.25080/Majora-92bf1922-00a.

24. F. Pedregosa et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–30; Radim Řehůřek and Petr Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta: ELRA, 2010), 45–50.

25. Michael Waskom et al., *Mwaskom/Seaborn: V0.8.1 (September 2017)*, version v0.8.1 (Zenodo, 2017), https://doi.org/10.5281/zenodo.883859.

26. Tom Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition, 27, no. 8 (June 1, 2006): 861–74, https://doi.org/10.1016/j.patrec.2005.10.010.

27. Walker, "Measuring the Impact That Domain Knowledge in the Form of Simple Ontology Structures Has on Predictive Modelling Processes in the Context of Academic Library Virtual Reference Services."

28. Python Software Foundation, "Re—Regular Expression Operations," Python 3.9.0 documentation, November 11, 2020, https://docs.python.org/3/library/re.html.

29. F. Pedregosa et al., "TfidfTransformer—Scikit-Learn 0.23.2 Documentation," scikit-learn: machine learning in Python (2020), https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html#sklearn.feature_extraction.text.TfidfTransformer.

30. Radim Řehůřek and Petr Sojka, "Models.Doc2vec – Doc2vec Paragraph Embeddings," gensim: topic modelling for humans, November 1, 2019, https://radimrehurek.com/gensim/models/doc2vec.html#gensim.models.doc2vec.

31. Le and Mikolov, "Distributed Representations of Sentences and Documents."

32. Walker, "Measuring the Impact That Domain Knowledge in the Form of Simple Ontology Structures Has on Predictive Modelling Processes in the Context of Academic Library Virtual Reference Services."

33. F. Pedregosa et al., "MLPClassifier," scikit-learn: machine learning in Python (2020), https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.