ICIS 2022 Proceedings

AI in Business and Society

Dec 12th, 12:00 AM

Development of an Automated Physician Review Classification System: A hybrid Machine Learning Approach

Sagarika Suresh THIMMANAYAKANAPALYA
university at buffalo, sthimman@Buffalo.edu

Pavankumar Mulgund
University at Buffalo, pmulgund@memphis.edu

Raj Sharman
University at Buffalo, SUNY, rsharman@buffalo.edu

Development of an Automated Physician Review Classification System: A Hybrid Machine Learning Approach

Completed Research Paper

Sagarika Suresh Thimmanayakanapalya
University at Buffalo
Buffalo, NY 14260, USA
sthimman@buffalo.edu

Pavankumar Mulgund
University of Memphis
Memphis, TN 38142, USA
pmulgund@memphis.edu

Raj Sharman
University at Buffalo
Buffalo, NY 14260, USA
rsharman@buffalo.edu

Abstract

Patients are increasingly turning to physician rating websites to help them make important healthcare decisions, such as selecting primary care doctors, specialists, and supplementary medical care providers. Previous research has identified a variety of topics and themes that emerge on these review platforms. However, there is little or no work that has been done to create an automated classifier that automatically categorizes these reviews into distinct topics after they have been explored in this context. Building such an automated classifier could assist IS developers and other stakeholders in automatically classifying patient reviews and understanding patient needs. Furthermore, using design science research we strategize how such machine learning systems can be built using design guidelines in turn having the potential to be generalized to other specific contextual problem spaces. Our work focuses on laying the foundation to design guidelines that need to be followed while building automated systems in specific contexts.

Keywords: Machine Learning, Design Science Research, Text Classifier, Physician Review Websites, Online Review Classification

Introduction

Physician rating websites (PRW) are patient-to-patient information sharing platforms where healthcare providers are rated and reviewed (Lagu et al., 2010). PRWs have witnessed a significant increase in their use and impact in the last few years. Increasingly, patients use PRWs to make critical healthcare decisions, including choosing primary care professionals, specialists, and ancillary medical care providers (Ko et al., 2019). Several studies have corroborated the increasing impact of PRWs on patient decision-making. For instance, a recent study noted that one in six Americans consult PRWs (Ko et al., 2019), and it is typically their first step in choosing providers (Frosch and Kaplan., 1999). Therefore, it is not surprising to find literature that demonstrates that patients not only select doctors based on positive ratings but also avoid doctors with poor ratings (Lu and Rui., 2018). PRWs are particularly popular with younger individuals aged 18 to 24 years, indicating that the popularity and use of PRWs are only slated to grow further (Galizzi et al., 2012). Moreover, PRWs are increasingly used by older adults, who constitute a significant segment of healthcare consumers. The COVID-19 pandemic has further enhanced the impact of PRWs, particularly for those patients seeking telemedicine services.

However, physician reviews and ratings have several limitations. First, patients may not be well-positioned to rate the physicians' services due to the information gap that exists between them. Second, patient reviews and ratings usually represent subjective opinions of their experiences and may not truly reflect the clinical quality of care provided by the doctors. A prior study points to the lack of reviews relating to the clinical aspects and notes that most of the review comments focus on non-clinical factors such as staff behaviors and appointment timelines (Lagu et al., 2010). However, as the use of PRWs increases, especially among middle-aged and older adults, reviews and ratings play a significant role in facilitating high-risk healthcare decisions such as choosing a specialist. Consequently, for this population segment, the clinical quality of care is a critical factor that supersedes all other elements.

Further, such a classification would be invaluable for patients of certain life-threatening or rare conditions where physician skill is the most critical success factor. Therefore, there is a need to classify physician reviews into clinical and non-clinical factors. Although such a classification offers considerable practical value, there is not enough previous work relating to designing and developing such automated classification systems. Additionally, such a classifier is not only helpful to other patients seeking information about physicians but also to other stakeholders such as providers, care quality managers, and practice/hospital administrators. Providers can mainly focus on the clinical aspects of reviews, and other stakeholders can address the non-clinical aspects. Moreover, these classifications can be very useful in identifying patterns in review comments as a first step in analyzing the root causes of the review comments. With the availability of current machine learning capabilities and a large volume of review data, self-learning models that create a hierarchical classification of review data into clinically relevant or process-relevant categories seem like a natural progression.

Some previous work has used text-mining methods to characterize patterns in physician reviews. For example, Wallace et al. (2014) created a probabilistic generative model to capture latent sentiment across many dimensions of care. Their work, however, emphasizes the emotion of the evaluations rather than collecting patterns and categorizing contextual aspects across them. Another work by Hao et al. compared reviews between Good Doctor Online and the US doctor review website RateMDs using topic modeling (Hao et al., 2017). While they discovered similarities between the two places, they also found variances representing the two countries' health care systems. Another work concentrated on building a classification model for physician recommender service based on the physician's information only, and it did not focus on the demands of the patient (Chiu & Cheng, 2016). Moreover, it utilized the dimensions based on physician knowledge from sources outside the physician rating website. (Li et al., 2018) built a mix-method approach to building a taxonomy for physician-related and patient-related domain attributes. However, they did not discuss building an automated classifier that can further classify new reviews. While most of the past literature focuses on highlighting themes and topics under physician rating websites, our paper advances the previous literature by proposing a design science research methodology of constructing a classifier by comparing three different neural network topologies and addressing specific lexical ambiguities to the PRW context. The novelty of this research lies in that it combines design science research with advanced machine learning methodologies as a foundation for future machine learning enthusiasts to innovate by using the design science process to improve proposed classification schemes further.

Our research project aims to design a machine learning model to help classify reviews into clinical and non-clinical categories following a design science approach (DSR). From a modeling perspective, we iteratively build hybrid machine learning models leveraging both supervised and unsupervised models to develop the classification. In this first iteration, we build a supervised model leveraging theoretical precursors to guide the classification process. In the second iteration, we use unsupervised learning approaches to refine the classification further to make it less ambiguous. In doing so, we demonstrate the application of DSR to the design and development of self-learning applications while highlighting methodological nuances relating to the development of these models.

Two objectives guide our subsequent efforts in this project: (a) designing and developing automated hybrid classification models to classify physician reviews by patients and (b) Reflecting on our learning experiences during the project to elaborate on the DSR process enhancements that cater to the nuances of designing self-learning applications. The remainder of the paper consists of explaining our research approach which consists of mainly three design cycles (i) the Relevance cycle which looks into the problem relevance, and the dataset relevance (ii) Model Development and Evaluation cycles which build and compare automated review labels with manually annotated data (iii) Model Refinement cycle and Expert Validation phase where we add further sub-topics for addressing misclassification of reviews and validate our main findings with experts. The paper ends with discussions and conclusions gathered from the findings.

Research Approach

Key Design Science Considerations

Although there is some literature on the development of automated classification systems using machine learning approaches, the focus has been on the development, training, and testing of the model with no emphasis on human involvement in the loop. We present a tabulation of prior efforts in Table 1

Paper	Description
Effectiveness of Fine-Tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews (Bilal and Almazroi, 2022)	Uses BERT to classify consumer reviews into helpful and unhelpful reviews using data Yelp. The authors compare the performance of Bert models against the bag of words. However, there is no manual annotation to develop training data or human-centered evaluation of the results. Moreover, the researchers do not point to the development strategies of datasets in the healthcare context
Text-mining-based fake news detection using ensemble methods (Reddy et al., 2022).	Use text mining and ensemble methods to classify news as either fake or real. The authors compare several models however the emphasis is no algorithmic performance
Combining Crowd and Machine Intelligence to Detect False News on Social Media (Wei et al, 2022)	Leverages a combination of crowd intelligence and unsupervised learning method to detect fake news. Uses human input to guide machine learning methods. In addition, uses unsupervised learning methods. However, there is no design science elements with a focus on human centered evaluation
Heart Disease Identification Method Using Machine Learning Classification (Ping Li et al., 2020)	Classified patients by identifying heart diseases using machine learning approach. However, there was design science approach or human involvement in assessing the application.

Table 1. Comparison to past literature

However, unlike these studies, our work weaved human input and human evaluation in the entire development and evaluation cycle. Therefore, we leveraged design science research (Hevner, 2004), which emphasizes problem relevance, design and evaluation cycles, and ongoing human-centered evaluation from end-users in line with the notion of stakeholder involvement outlined by Sein et al. (2011). However, our efforts in this research highlight several interesting issues that emerge when the design science approach is utilized to develop machine learning/AI applications. The following are some fundamental orientations that differentiate our research effort from the traditional machine learning work.

First, much of the prior work (Hevner, 2007; Drechsler & Hevner, 2018) in DSR has pointed to three cycles: 1. Relevance 2. Design and 3. Rigor. While the relevance cycle focuses on "exploring the application context that provides requirements and defines acceptance criteria" (Hevner, 2007), the Design cycle addresses iterative development and evaluation of the design artifacts. The rigor cycle emphasizes additions of knowledge base theoretical contributions. Although the three-cycle model of DSR and its extensions are an excellent fit for the design of traditional IS research artifacts such as mobile apps and web applications, there is not enough literature on some of the nuances that self-learning applications such as AI and ML applications bring to the fore.

For instance, in the case of AI and ML applications, the relevance of the problem is established not only by the external context, such as user needs but also by the data used to train and test these applications. For instance, the functioning of a recommendation engine is influenced far more significantly by the data used to train the model than any external requirements that end-users specify. Although end users may point to specific metrics that act as the north star, it is usually the dataset that decides specific behaviors of self-learning applications. Consequently, there is a need to accommodate the use of prior data as a source of the requirement for self-learning applications. Specifically, tactical guidance on the choice of apt external data sources, data validation in terms of data-audience fit, continuous adjustment (improvement) of data, and the use of prior theoretical work to guide the identification of key labels is needed. We specifically use manual annotation of data using qualitative coding using prior theoretical work as training data, which helps guide the behavior of the automated classification system. In doing so, we demonstrate how theory guides qualitative thematic analysis and can be used as a preprocessing step to develop machine learning models

Second, our work followed an iterative approach with multiple design-evaluate cycles. Each design-evaluate cycle consisted of formative evaluation (Venable, 2016) followed by a redesign effort, based on the suggestions from Hevner et al. We position the training, testing, and benchmarking models as design and evaluate cycles because the best classification system evolves from not only developing different machine learning models but also iteratively adjusting the parameters within the model. The outcome of our work —an automated classification system— itself represents a machine learning-based design artifact. Due to this design orientation. Our work uses a combination of supervised and unsupervised approaches to build and refine the classification respectively.

Third, our evaluation efforts include the evaluation of design science outcomes (Venable, 2016) as well as the evaluation of machine learning models (Flach, 2019). From the machine learning perspective, we continually assess the model for accuracy and precision using model fit statistics such as ROC curve and confusion matrix using test data. We also benchmark different models to choose the one that best performs. Further, from a design science perspective, we perform continuous assessments of problems, data, and design artifacts. Finally, our efforts concluded with a summative human-centered assessment of the classification algorithm that compared human classification with the results of automated classification. The human-centered assessment has been mostly ignored by prior work. Therefore, this work highlights a critical issue in the evaluation of machine learning artifacts.

Fourth, although this work describes a specific instance of automatically classifying patient reviews of physicians into clinically relevant or irrelevant ones, this work addresses a general problem of automatic classification of consumer reviews about specialized services into different predefined categories. Table 2. summarizes the critical components of this research.

Research Component	**Description**
Practice inspired problem	Points to a class of problems that can be addressed by the development of classification models for user generated content
Design and evaluate iterations	Consists of training and testing machine learning models as design-evaluate cycles (Hevner et al., 2004) with formative evaluation and refining of models
Use of a prior research as kernel theory	Uses prior review study (Rothenfluh & Schulz, 2018). as initial classification that guides our supervised learning models
Continuous assessment and improvement	Identifies several assessment and ongoing improvements to the machine learning model from both process and outcome perspective

Table 2. Design Science Research Components

Overall Design Process

In this paper, we use three design iterations. In the first design iteration, we establish the need and relevance of developing the classification system. We draw from previous literature that points to the need for a physician needs classification system (Segal et al., 2012). From a design science perspective, we emphasize the salient differences that exist between traditional IT applications and self-learning applications such as classification systems. Unlike traditional design artifacts such as web and mobile applications, the relevance in self-learning applications is as much established by the data used to train the model as it is by the context and the end-user needs (Zaknich, A., 2005). Therefore, we argue that activities such as data source identification for training, data cleansing, and preprocessing are critical to establishing relevance. We also note that the choice of the data source depends on the context and end-user needs.

Consequently, as presented in Figure 1, activities such as data source identification, data cleansing, and data preprocessing are positioned as relevance-related activities performed after understanding end-user goals and context. Further, we also point to the ongoing nature of these activities and a need for continuous improvement because as the context changes, the data necessary to train the models must also change. For instance, in a recommendation engine, the training data should evolve with changing demography and tastes of the customers.

In the second iteration, we focus on design and evaluation cycles. The design process involves choosing the relevant machine learning models and formatively assessing and benchmarking models using model fit statistics. The emphasis is on evaluating the model from a machine learning perspective and fine-tuning the parameters of the models to improve classification accuracy. In addition, we benchmark one model against another. We use LSTM, BI-LSTM, and BERT to classify the data. Following that, we compare the model results to manually annotated data sets to evaluate the models further. Using the results from these assessments, we refine the classification through the process of ambiguous class disintegration.

In the third iteration, we generate sub-classifications leveraging unsupervised learning models such as LDA. Following that, we perform some summative tests involving actual end-users to obtain their perspectives

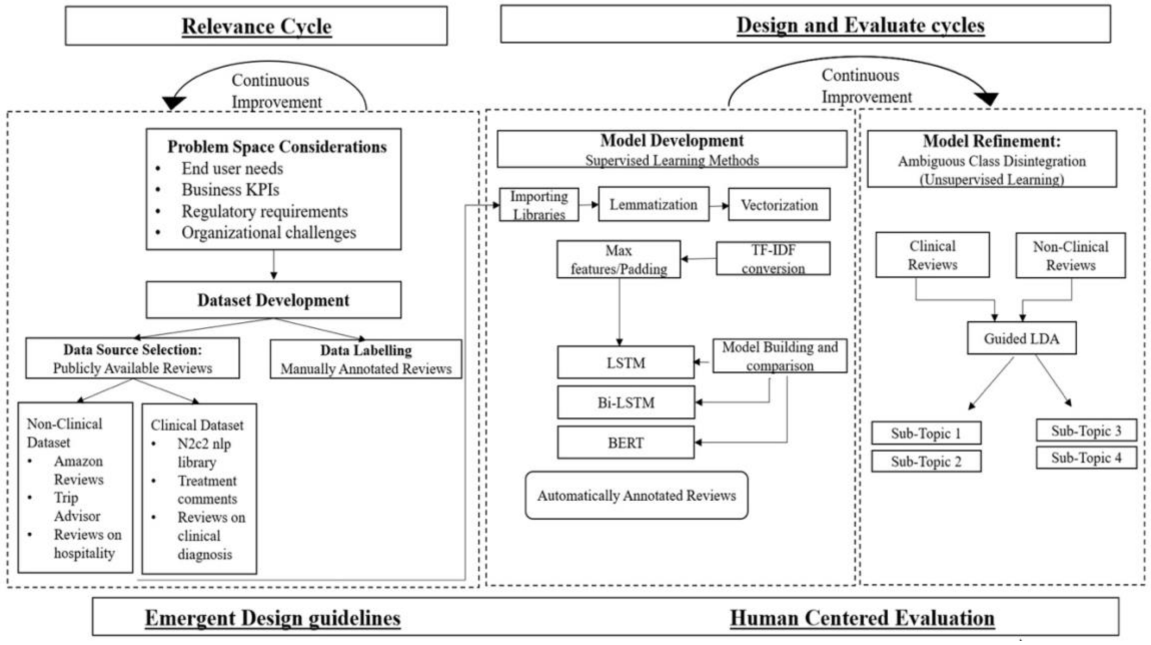

Figure 1. Modified Research Model

Relevance Cycle

In the relevance cycle, there were two types of datasets used for our methodology. One was the manually annotated data, and the other was the tagged corpora based on widely available clinical and non-clinical datasets. In this section, we explore the difficulty in the process of obtaining and labeling datasets for building an automated classifier. Moreover, we explain the design of cleaning the datasets once they are obtained for using them as training data in our classification models.

Manual Annotator Labeling

We used the manually annotated data to understand how physician reviews were broadly segregated. Using earlier work by Tavakol et al. (2006) for grounded theory in medical education research, the grounded theory investigators and manual annotators found that reviews in RateMD could be classified into two main topics, which included clinical and non-clinical reviews. This was dataset exploration and understanding broadly how physician reviews could potentially be classified. To develop this dataset, we collected datasets from a major physician review website (PRW), which consisted of fundamental physician ratings. Three researchers qualitatively coded the data using the guidelines provided by the grounded theory method (Charmaz., 2008). Each of the ratings was classified as either clinical or non-clinical based on the metrics manual for the coding developed and curated from previous literature (Tavakol et al., 2006). However, when a review comment consisted of both clinical and non-clinical parts, the coders were asked to subjectively assess which theme was dominant to classify the review. The interrater agreement reliability score was computed after the coding and interrater reliability of 0.79 was observed. When the coders differed in their classification of the review comment, a majority vote was used to finalize the classification. In some cases, when there was no consensus, an expert qualitative coder was consulted to finalize the classification. Table 3 below allows us to understand how the classification between clinical and non-clinical reviews was conducted. More metrics mapping was adopted from Rothenfluh & Schulz (2018).

Clinical	Non-clinical
Reviews related to treatment procedures	Review related to appointment scheduling
Reviews related to medication provided by doctor	Review related to staff hospitality
Reviews related to doctor’s knowledge	Review related to hospital atmosphere

Table 3. Sample of how reviews were coded for ground truth

Labeled Corpora for Supervised Learning

The manually annotated dataset allowed researchers to explore the main requirements from comments on online physician review sites. After reviewing prior literature and exploring the problem, we acquired public datasets. Prior research has not attempted to develop classifiers that segregate and label data into clinical and non-clinical data, possibly because of the obscurity of datasets in the healthcare field due to compliance issues. Therefore, we curated the dataset by collecting and combining data from various verified sources. From a machine learning perspective, making a machine understand non-clinical data involved training the algorithm with sentences adapted from reviews on trips, hotels, airports, movies, books, and various general reviews on appointment scheduling at a lawyer’s office. This allowed our model to understand non-clinical generic vocabulary.

The clinical data was comparatively harder to obtain because of compliance and data verification issues. However, we secured freely available datasets containing medical transcription samples. The data was originally produced during the i2b2 project (Informatics for Integrating Biology & the Bedside - A National Center for Biomedical Computing), and it consisted of clinical review notes from the n2c2 NLP research data sets (Harvard medical school) (Oleynik et al., 2019). This dataset comprised sample medical transcriptions for various medical disciplines, allowing us to arrange statements like "diagnostic about a heart disease was well done" and "treatment was done in great detail" as clinical reviews. Figure 2 provides the dataset exploration process. The n2c2 NLP dataset allowed the machine learning classifier to understand the wording complexities of the medical field. Consequently, we had two different datasets to improve our classification. While the publicly available dataset (labeled corpora) was used for the classification development process, and the manually annotated labels were used to check the misclassification rate and validate the different types of classifications built.

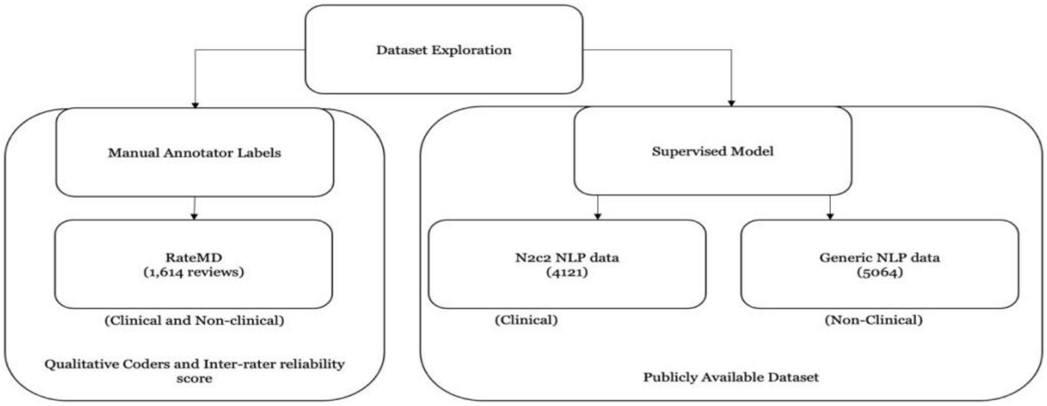

Figure 2. Dataset Information

Text Pre-processing

Before developing the classifier, we conducted all the necessary pre-processing steps on the labeled corpora. Textual data, such as reviews, is different from numerical data, and such data is represented in human language and is not easy to convert directly into the quantitative format. In addition, processing raw text directly could be very noisy because some text content may not contain useful information. We used a natural language processing toolkit (NLTK) to process our data (Loper and Bird 2002). Detailed processing steps are shown in Figure 3. Numbers, punctuations, and stop words were removed and converted to lowercase for uniformity. The texts were then stemmed and lemmatized.

Since the machines work optimally with an array of numbers instead of a set of strings while training the classification models, we converted the strings into TF-IDF format (text to numbers). We then extracted the max features (Aizawa, 2003). Following this, padding was done to extract meaningful features from the context of the corpora presented (Dwarampudi & Reddy, 2019). Out of the total corpora, the dataset was imbalanced with 5064 clinical texts, and 4121 non-clinical were selected according to the standards of previous literature (Jain et al., 2013). The training data was split as 80% of the collected data points, and validation was 20% of the collected data points. The manually annotated data (from RateMD) (Figure 2) was used as the test data, which consisted of 1,614 reviews. For our machine learning models, the number of epochs was set to 11, which was optimally decided based on our dataset size and the number of classification labels. The batch size was set to 32 samples because we wanted to keep the training samples smaller per epoch to yield optimal results (Jain et al., 2013).

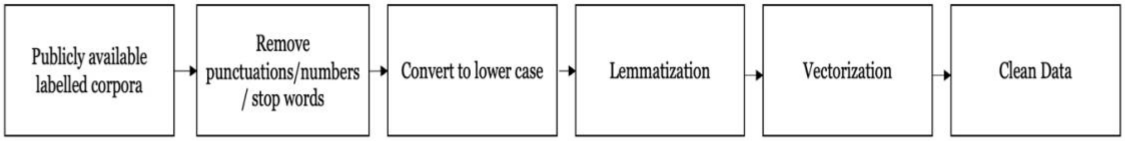
Publicly available labelled corpora
Remove punctuations/numbers / stop words
Convert to lower case
Lemmatization
Vectorization
Clean Data

Figure 3. Dataset Preprocessing

Design and Evaluate Cycles

In this sub-section, we will describe how the models were developed. Figure 1 provides a bigger picture of how the design and evaluation cycles of the model design. Using the labeled corpora from publicly available datasets, we trained and built three types of supervised classification models: the Long Short-Term Memory network (LSTM), Bi-directional Long Short Term Memory Network (Bi-LSTM), and the BERT Bidirectional encoder representations from transformers. We chose these three models to understand which classification models accurately classify healthcare reviews. Moreover, comparing three different classification models allowed us to test and evaluate the best classification framework repeatedly. Furthermore, we also chose BERT over ROBERTa, DistilBERT, XLNet because DistilBERT is known to produce 3% degradation from BERT. Although ROBERTa, and XLNet give better performance measures, their training time is high, and the data size required is higher (Cortiz, 2021; Adoma et al., 2020).

Model Development using Supervised Learning

We tested three main models, including deep neural network learning models that followed supervised learning mechanisms. The models were (1) Long short-term memory (LSTM) neural networks, (2) Bidirectional -LSTM, and (3) Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). We will describe each model-building process separately and compare the three model classifiers in the results section.

Supervised Learning (LSTM)

We first used the long short-term memory network (LSTM) to build the classifier. Long Short-term memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. Recurrent neural networks are different from traditional feed-forward neural networks (Sundermeyer et al., 2012). The additional complexity comes with the promise of new behaviors that conventional methods cannot achieve. Recurrent neural networks have an internal state that can represent context information, the critical information about the past inputs for an amount of time that is not fixed a priori but depends on its weights and the input data.

A recurrent neural network whose inputs are not fixed but constitute an input sequence can be used to transform an input sequence into an output sequence while taking contextual information flexibly (Lipton et al., 2015). The very reason for selecting a long-short-term memory neural network is that our goal is to build a classifier that adapts to new contexts. Choosing an LSTM model helped us to account for context adaptability additionally. We fit the detection model using LSTM. Some new hyper-parameters used in LSTM were *the* number of nodes in the hidden layers, which we chose to be 20 within the LSTM cell, and also the actual value set for return sequences ensures that the LSTM cell returns all of the outputs from the unrolled LSTM cell through time. If this argument is not used, the LSTM cell will provide the result of the LSTM cell from the previous step.

Supervised Learning (Bi-LSTM)

Next, we used the bidirectional LSTM (Bi-LSTM) networks to build our classifier. Unlike LSTM, the Bi-LSTM learns patterns from both before and after a given token within a document (Zhang et al., 2020). The Bi-LSTM back-propagates in both backward and forward directions in time. Due to this, the computational time was increased compared to LSTM. However, in most cases, Bi-LSTM was said to result in better accuracy. We expected the model to perform well overall through the Bi-LSTM model because building this model helped us further investigate whether adding backward and forward propagation towards our curated datasets helped improve its context adaptability (Mughees et al., 2021).

Bidirectional Encoder Representations from Transformers (BERT)

Finally, we used BERT classification to build our classifier. BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Gardner et al., 2018; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both the left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications (Devlin et al., 2018). The main reason we decided to use BERT is to address the lexical ambiguity specifically concerning the healthcare context, and words embedded in our context especially. The results and comparisons of the three models are given in the next subsection.

Binary Classification Results

The results of the three models can be seen in Table 4. The BERT classification model reported an accuracy of 99.7%, the LSTM network model reported an accuracy of 99%, and the Bi-LSTM network reported an accuracy of 99.3%. We also compared the precision, f-measure, and recall of the three models. The BERT algorithm performed optimally in terms of these measures.

Measure	BERT	Bi-LSTM	LSTM
Accuracy	0.997	0.993	0.990
Precision	0.948	0.869	0.894
F-Measure	0.865	0.564	0.279
Recall	0.795	0.417	0.165

Table 4. Comparison of Supervised Machine Learning Models

Validation and Comparison of the Models with the Manually Annotated Data

In this subsection, we compared the output from the three classification models and realized the misclassification rate of reviews. An output score closer to 1 was a non-clinical review, and an output closer to 0 was a review clinical in nature. In table 5, we compare the model predictions which used labeled supervised corpora with the manually annotated labeling and notice that in both clinical and non-clinical classification, the BERT model does the best job at classifying the reviews with a predicted score of 0.99. In review 1, which was non-clinical and purely about the expense, the Bi-Directional model predicted a relatively safe value of 0.98. The LSTM model scored 0.66 which was not as accurate as the other two models. Similarly, BERT classified the clinical reviews as the best with a prediction score of 0.0078 which was closer to 0.

Sl No.	Review	Model type	Model prediction 'clinical': 0, 'nonclinical': 1	Manual Classification Label (Ground Truth)
1.	"Expensive doctor, too much bills to pay"	LSTM	0.66	Non-clinical
	"Expensive doctor, too much bills to pay"	Bi-LSTM	0.98	Non-clinical
	"Expensive doctor, too much bills to pay"	BERT	0.99	Non-clinical
2.	"Dr Ahmed took care of my mom when rushed to the hosp cuz she was septic and had no pressure.. He figured out what was wrong with her and saved her life. He was good at explaining things and answering questions I had. "	LSTM	0.13	Clinical
	"Dr Ahmed took care of my mom when rushed to the hosp cuz she was septic and had no pressure. He figured out what was wrong with her and saved her life. He was good at explaining things and answering questions I had. "	Bi-LSTM	0.24	Clinical
	"Dr Ahmed took care of my mom when rushed to the hosp cuz she was septic and had no pressure.. He figured out what was wrong with her and saved her life. He was good at explaining things and answering questions I had. "	BERT	0.0078	Clinical

Table 5. Comparisons of the review comments with the ground truth.

The BERT is mainly a model used to address lexical ambiguity, which is the greatest challenge to natural language understanding. The reason BERT was able to classify our analogy the best is because it is deeply bidirectional. Bidirectional means that BERT learns information from both the left and the right sides of a token's context during the training phase (Devlin et al., 2018). For instance, in our case, the word "treatment" could be used in a clinical as well as a non-clinical context. Clinically, it could be used as "The Doctor's treatment included long surgery hours." whereas non-clinically, it could be used as, "The staff treated me well and was friendly." The word treat here is used in two different contexts. BERT helps capture the left and right context, such as in the clinical statement above. BERT picks up contexts such as "doctor" and "surgery" and classifies it as a clinical review. The repeated design and validation across automated and manually annotated labeling allowed us to understand the complexities our specific dataset was challenging us with, such as contextual embeddings or words like "treatment." We, therefore, moved on to our final design cycle of model refinement to understand how to address complexities such as lexical ambiguity concerning our context.

Model Refinement Cycle

Ambiguous Class Disintegration

Cross-validating automated labeling with the manually annotated labeling in our design and evaluate cycles (Table 5.) across 1,614 reviews; we noticed that 1,311 reviews matched exactly in comparison. However, there were 303 misclassified reviews. This meant that although the BERT classification was used, it still accounted for 81% accuracy of correctly classified reviews. This showed us that even though the hyperparameters, learning, and validation score were good methodologically, there was still a hindrance to correctly classifying the reviews. Our confidence interval was therefore set to 81% and the remaining 19% as an ambiguous class which our algorithm did not determine. In this section, we, therefore, realize that the binary classification had to be further expanded into multi-class labeling to enhance the meanings and context of the dataset. Our limited dataset was manually classified as clinical and non-clinical for supervised learning algorithms. However, to further establish multi-classes, there was no availability of multi-class labeled data further than clinical and non-clinical. Therefore, we decided to establish multi-classes through a semi-supervised guided Latent Dirichlet allocation algorithm (Toubia et al., 2019). This would allow us to understand if there are multiple classes present under each of the clinical and non-clinical reviews, and the scores majorly fell from 0 to 1.

Subtopic Generation using Guided LDA

Previously, we noticed how many misclassified reviews fell between the prediction score of 0.4 and 0.8 using the BERT algorithm. We hypothesize in this subsection why the BERT allocation cannot work well in addressing some lexical ambiguities because the class still needs further disintegration. To understand if sub-topic generation caused the misclassification, we used Guided Latent Dirichlet Allocation to classify each non-clinical and clinical review into more granular classes. Guided LDA or SeededLDA implements latent Dirichlet allocation (LDA) using collapsed Gibb’s sampling (Toubia et al., 2019). GuidedLDA can be guided by setting some seed words per topic, making topics converge in a set direction. In our case, a good analysis of the grounding of the physician reviews and metrics manual was used to decide the seed words (Chapman et al., 2015). In Table 6, we see the seed words and their topic labeling; these match the metrics manual developed by us.

Clinical seed words	Non-clinical seed words
'knowledge','competance','correctness','diagnostic','ability','timely','referral','completeness’, ‘quality','cost','consciousness', 'testing','experience','responsible','systematic','correct','quality'	'environment','cleanliness','comfort','instrument','execution','treatment','procedure','reachability','punctuality','scheduling','waiting','time','notification','reachability','notification','appointment','teamwork','staff','monitoring','training','provisioning','comprehensiveness','social','skills','attentiveness','privacy','protection','shared','decision','communication','recommendation''satisfaction','efficiency','complication','follow-up'

Table 6. Seed words for the Guided LDA model from the metrics manual

Once the clinical and non-clinical classes were established, we further analyzed whether each review could be split into more granular level topics. Table 7 shows that clinical reviews could be split into five subclasses physician knowledge, diagnosis, treatment, physician demeanor, and clinical staff treatment. Non-clinical reviews were split into five sub-classes: appointment scheduling, bedside manner, relationship building, staff professionalism, and patient experience. The rationale for classifying each of them under these different classes is given in Table 7 under rationale. We checked that the coherence and perplexity scores for the clinical review sub-classification were 0.5 and -0.6, and non-clinical were 0.48 and -0.56, each of which had optimal scores. Moreover, we determined the automated score range of the different subtopics by comparing the number of correctly classified reviews (where automated and manual scores matched) with the automated score range that they fell under. For instance, 98% of correctly classified reviews under the sub-topic "physician knowledge" fell between the closed range of $0 \leq x \leq 0.1$ or [0,0.1], as shown in Table 7.

Furthermore, we also used the InfraNodus to calculate if the discourse on each of these topics was focused on one main topic (Paranyushkin, 2019). For instance, InfraNodus allowed us to understand if the topics diverged in a review or pointed towards one specific sub-topic. This allowed us to understand how much each review was consolidated within a particular sub-topic. For example, the example review under the sub-topic "physician review" highly focuses on the physician's knowledge and, therefore, is classified as a focused discourse in InfraNodus.

Multi-Class Labeling	**Automated Score Range**	**Example Review**	**Rationale**
Clinical [0]			
Physician Knowledge	[0, 0.1]	"I left appointment Dr Kevin Gibbons crying hard I developed nosebleed. Dr Gibbons may knowledgeable aspects field, I devastated find dismissive connection neurological issues rare disease - even worse questioning diagnosis rare disease (telling rare unlikely I actually it)."	Reviews spoke about physician's knowledge about the treatment, surgery
Diagnosis	[0.1, 0.2]	"I went see Dr Gibbons Cervical disk damage bothering years. After full series X-rays, MRIs, etc., diagnosed condition explained advocates conservative approach treatment surgery last option. I appreciated honesty concern even though meant less business him. I definitely see Dr Gibbons condition progresses point I consider surgery."	Reviews spoke about how the diagnosis procedure was and whether the physician conducted the right diagnosis
Treatment	[0.2, 0.3]	"Exceptional doctor seems genuinely care patient extremely thorough developing staged treatment plan looking solutions outcomes condition. Very knowledgeable (especially regard asthma) share insights explanations clear understandable manner."	Reviews spoke about the treatment procedure, and whether the treatment was thorough

Physician Demeanor	[0.3, 0.4]	"Was patient empathetic re:: problem. Took time locate source pain treat. Also made recommendations future self-treatment. Would recommend others."	Reviews spoke about the physician's demeanor during clinical procedures, and treatments
Staff Clinical Treatment	[0.4, 0.6]	"All one receptionists, offices, nasty rude. They perform physicals week. They chronic disease farm. He interested seeing chronic patients need see twice year, $75+ per visit, write reoccurring prescriptions. "	Reviews spoke about those diseases where staff prescription, and clinical help from nurses was required
Non-Clinical [1]			
Appointment Scheduling	[0.4, 0.5]	"I felt bit rushed, made sure I got questions answered I left, hence satisfactory rating I gave above. This another doctors office seems book patients handle efficiently. The patients end waiting hour seen."	Reviews spoke about the ease of appointment scheduling process and their experience with it
Bedside Manner	[0.6, 0.7]	"Good doctor, bedside manners: fixed problem: thank "	Reviews spoke about bedside manners
Relationship Building	[0.7, 0.8]	"Dr. Pagano sense humor. Makes every visit calming & easy. "	Reviews spoke about the relationship building etiquettes between patient provider
Staff Professionalism	[0.8, 0.9]	"Professional staff really cares being."	Reviews spoke about the professionalism of the staff
Patient Experience	[0.8, 0.9]	"Overall pleasant experience visit Alison Graziano. "	Reviews spoke about patient overall experience.

Table 7. Subtopic Generation Strategy for Model Refinement

Similar to Table 7, we also checked the sub-classes that occurred under mismatched reviews i.e., the mismatch between clinical and non-clinical reviews. While the matched reviews allowed us to understand the different multi-class labeling, we can develop in the future, the mismatched reviews allowed us to understand various lexical ambiguities that may have caused this misclassification. Our further exploration of sub-classes under mismatched reviews showed us that money issues related to **ethical behavior of physicians** and staff overall, **emergency room visits** which contained mixed reviews on the staff, and physician, **prescription-based reviews** which again required the involvement of physician and staff interoperability mostly contributed toward misclassification. This is because, while most reviews in correctly classified reviews consisted of discourse scores that were in the same focus (either focused on non-clinical or clinical), the misclassified reviews had dispersed discourse in reviews (consisted of topics that would require the involvement of both clinical and non-clinical aspects. The focus of our paper was to explain and understand the reasons behind ambiguity in the BERT model due to dispersed discourse scores (containing a mixture of clinical and non-clinical discourse aspects). However, there is limited publicly available data that classified clinical and non-clinical reviews into further subclasses, and therefore, we furthered this study by analyzing what subgroups and topics are online physician reviews about and have laid a foundation for future machine learning enthusiasts in the healthcare domain to improve publicly available datasets concerning the above-mentioned subtopics. Furthermore, these sub-topics also give us an understanding, of the kind of reviews, and expectations patients have from physicians. The model refinement cycle, therefore, pointed out how the model can further be improved in terms of generating multi-level classes, and creating more labels for supervised learning. Building such datasets with more interpretable sub-topic labeling in the future would help build much more accurate classification systems.

Expert Validation

In addition to ongoing formative analysis, we evaluate the outcome of the model through expert evaluation. Our evaluation was based on the feedback from machine learning experts from industry and academia. We invited eight experts five from academia (3 design science and 2 machine learning researchers) and three from industry to assess the classification model. Industry practitioners assessed the model using sample testing (Vabalas et al., 2019), a quick and effective method of assessing classification. In this method, the practitioners as a group selected a sample of reviews and manually classified the reviews into certain categories. When there were differences among them, they reconciled the differences through deliberation, but they marked the review as high risk for misclassification. Following that, the experts input the same review comments as input in the classification system. They compared the machine output against humans to assess the model. The results were very promising with 99.37% accuracy. The academic experts reviewed the process used to develop the classification system. We adopted the design walkthrough approach (Hermann., 2009) to validate the process. We systemically demonstrated the design science approach, data collection and cleaning process, choice of models and our results, and training and testing results. Following that, we elaborated on the hybrid approach. Based on the walkthrough several improvements were suggested. Table 8 below summarizes the findings

Data related aspects - Data from other patient rating websites such as vitals.com, health grades and Zocdoc may reveal interesting insights - Determine data sufficiency through benchmarking models running with different volumes of data
Present the confidence level in the models - Classified models must also have an indicator of confidence making it easier for user to spot some of the lexical ambiguities during classifications
Improve hybridization - Need to compare the results of unsupervised model to hybrid models for additional benchmarking
Design science process related issues - Process to identify the need for hybridization - Elaborate on the disambiguation process - Potentially describe lean approaches to designing AI/ML applications - Emphasize on design principles such as transparency, interpretability in ML systems

Table 8. Feedback from experts

Discussion and Conclusion

In this paper, we pursued two goals: (a) developing a physician review classification system and (b) Elaborating on a design science approach for developing a self-learning application. These goals reflect salient issues for the design of next-generation applications. Much of prior work in design science has emphasized traditional IS applications such as mobile phones and web applications. Further, prior work in the machine learning area on physician reviews by patients has focused on generating keywords and themes, but there is not enough work on classifying the physician reviews as clinically relevant or not. Our effort in this paper, therefore, attempts to build such a classification model and further refine it using hybrid approaches. Our work stands on two strands: (a) machine learning model development and (b) design science process to develop the classification system.

In doing so, we have attempted to make several research contributions. First, we elaborate on a design science research approach to develop a self-learning application. In this process, we detail the relevance, and design and evaluate cycles as it applies to self-learning applications. Second, from a machine learning perspective, we develop a hybrid classification model. In the initial iteration, we develop a supervised classification model and test it with manually classified data. In a later iteration, we perform ambiguous class integration to refine the classification model. Following that, we generate subclasses that utilize the hybrid mechanism to assign estimated automated score ranges. Moreover, with the introduction of subclasses, we invite future researchers to develop labeled subclasses to improve machine learning models specific to the PRW context. The hybridization of models is unique in that it not just classifies based on existing categories but generates new ones as new data is accumulated.

There are some limitations to our work. First, there are several aspects of the DSR process that need to be elaborated on as identified in the expert review. Second, our expert evaluation was based on the feedback of a small sample of eight experts. Another limitation is that our work has not elaborated on the design principles as it relates to the development of self-learning application. The attempt in this study has been to apply the DSR approach to develop a classification model. Our future agenda for this work will be to reflect and formalize design principles and evaluate those principles for generalizability.

References

Ackoff, R. L. 1961. "Management Misinformation Systems," Management Science (14:4), pp. 147-156.

Adoma, A. F., Henry, N. M., & Chen, W. (2020, December). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 117-121). IEEE.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45-65.

Benbasat, I., and Zmud, R. W. 2003. "The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties," MIS Quarterly (27:2), pp. 183-194.

Bilal, M., & Almazroi, A. A. (2022). Effectiveness of Fine-Tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews. Electronic Commerce Research, 1-21.

Bonini, C. P. 1963. Simulation of Information and Decision Systems in the Firm, Englewood Cliffs, NJ: Prentice-Hall.

Broadbent, M., Weill, P., O'Brien, T., and Neo, B. S. 1996. "Firm Context and Patterns of IT Infrastructure Capability," in Proceedings of the 14th International Conference on Information Systems, J. I. DeGross, S. Jarvenpaa, and A. Srinivasan (eds.), Cleveland, OH, pp. 174-194.

Carroll, J. 2005. "The Blacksburgh Electronic Village: A Study in Community Computing," in Digitial Cities III: Information Technologies for Social Capital, P. van den Besselaar and S. Kiozumi (eds.), New York: Springer-Verlag, pp. 43-65.

Chapman, A. L., Hadfield, M., & Chapman, C. J. (2015). Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. Journal of the Royal College of Physicians of Edinburgh, 45(3), 201-205.

Charmaz, K. (2008). Grounded theory as an emergent method. Handbook of emergent methods, 155, 172.

Chiu, M. H., & Cheng, W. C. (2016, July). Building a classification model for physician recommender service based on needs for physician information. In International Conference on HCI in Business, Government, and Organizations (pp. 28-38). Springer, Cham.

Cortiz, D. (2021). Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. arXiv preprint arXiv:2104.02041.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Drechsler, A., & Hevner, A. R. (2018, June). Utilizing, producing, and contributing design knowledge in DSR projects. In International Conference on Design Science Research in Information Systems and Technology (pp. 82-97).Springer, Cham.

Dwarampudi, M., & Reddy, N. V. (2019). Effects of padding on LSTMs and CNNs. arXiv preprint arXiv:1903.07288.

Flach, P. (2019, July). Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9808-9814).

Frosch, D. L., & Kaplan, R. M. (1999). Shared decision making in clinical medicine: past research and future directions. American journal of preventive medicine, 17(4), 285-294.

Galizzi, M. M., Miraldo, M., Stavropoulou, C., Desai, M., Jayatunga, W., Joshi, M., & Parikh, S. (2012). Who is more likely to use doctor-rating websites, and why? A cross-sectional study in London. BMJ open, 2(6), e001493.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., ... & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640.

Hao, H., Zhang, K., Wang, W., & Gao, G. (2017). A tale of two countries: International comparison of online doctor reviews between China and the United States. International journal of medical informatics, 99, 37-44.

Herrmann, T. (2009). Systems design with the socio-technical walkthrough. In Handbook of research on socio-technical design and social networking systems (pp. 336-351). IGI Global.

Hevner, A. R. (2007). A three cycle view of design science research. Scandinavian journal of information systems, 19(2), 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 75-105.

Jain, N., Bhatele, A., Robson, M. P., Gamblin, T., & Kale, L. V. (2013, November). Predicting application performance using supervised learning on communication features. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (pp. 1-12).

Jørgensen, M., & Papatheocharous, E. (2015, August). Believing is seeing: Confirmation bias studies in software engineering. In 2015 41st Euromicro Conference on Software Engineering and Advanced Applications (pp. 92-95). IEEE.

Ko, D. G., Mai, F., Shan, Z., & Zhang, D. (2019). Operational efficiency and patient-centered health care: A view from online physician reviews. Journal of Operations Management, 65(4), 353-379.

Lagu, T., Hannon, N. S., Rothberg, M. B., & Lindenauer, P. K. (2010). Patients' evaluations of health care providers in the era of social networking: an analysis of physician-rating websites. Journal of general internal medicine, 25(9), 942-946.

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. IEEE Access, 8, 107562-107582.

Li, J., Liu, M., Li, X., Liu, X., & Liu, J. (2018). Developing embedded taxonomy and mining patients' interests from web-based physician reviews: mixed-methods approach. Journal of medical Internet research, 20(8), e8868.

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2015). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. arXiv preprint cs/0205028

Lu, S. F., & Rui, H. (2018). Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. Management Science, 64(6), 2557-2573.

Mughees, N., Mohsin, S. A., Mughees, A., & Mughees, A. (2021). Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting. Expert Systems with Applications, 175, 114844.

Oleynik, M., Kugic, A., Kasáč, Z., & Kreuzthaler, M. (2019). Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. Journal of the American Medical Informatics Association, 26(11), 1247-1254.

Paranyushkin, D. (2019, May). InfraNodus: Generating insight using text network analysis. In The world wide web conference (pp. 3584-3589)

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Reddy, H., Raj, N., Gala, M., & Basava, A. (2020). Text-mining-based fake news detection using ensemble methods. International Journal of Automation and Computing, 17(2), 210-221.

Rothenfluh, F., & Schulz, P. J. (2018). Content, quality, and assessment tools of physician-rating websites in 12 countries: quantitative analysis. Journal of medical Internet research, 20(6), e9105.

Segal, J., Sacopulos, M., Sheets, V., Thurston, I., Brooks, K., & Puccia, R. (2012). Online doctor reviews: do they track surgeon volume, a proxy for quality of care?. Journal of medical Internet research, 14(2), e2005.

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action design research. MIS quarterly, 37-56.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In Thirteenth annual conference of the international speech communication association.

Tavakol, M., Torabi, S., & Akbar Zeinaloo, A. (2006). Grounded theory in medical education research. Medical Education Online, 11(1), 4607.

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent Dirichlet allocation approach informed by the psychology of media consumption. Journal of Marketing Research, 56(1), 18-36.

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. Journal of Marketing Research, 56(1), 18-36.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with limited sample size. PloS one, 14(11), e0224365.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. European journal of information systems, 25(1), 77-89.

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. Journal of the American Medical Informatics Association, 21(6), 1098-1103.

Wei, X., Zhang, Z., Zhang, M., Chen, W., & Zeng, D. D. (2019). Combining Crowd and Machine Intelligence to Detect False News on Social Media. MIS Quarterly.

Zaknich, A. (2005). Principles of adaptive filters and self-learning systems. Springer Science & Business Media.

Zhang, B., Zhang, H., Zhao, G., & Lian, J. (2020). Constructing a PM2. 5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. Environmental Modelling & Software, 124, 104600.