International Journal of Research in Marketing 39 (2022) 1–19

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full length article

An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews

Huwail J. Alantari [a], Imran S. Currim [a], Yiting Deng [b], Sameer Singh [c]

[a] Paul Merage School of Business, University of California, Irvine, United States
[b] UCL School of Management, University College London, London, United Kingdom
[c] Donald Bren School of Information and Computer Sciences, University of California, Irvine, United States

ARTICLE INFO

Article history:
First received on July 7, 2020 and was under review for 9 months
Available online 9 November 2021

Area Editor: Bernd Skiera

Keywords:
Automated text analysis
Sentiment analysis
Online reviews
User generated content
Machine learning
Natural language processing

ABSTRACT

The amount of digital text-based consumer review data has increased dramatically and there exist many machine learning approaches for automated text-based sentiment analysis. Marketing researchers have employed various methods for analyzing text reviews but lack a comprehensive comparison of their performance to guide method selection in future applications. We focus on the fundamental relationship between a consumer's overall empirical evaluation, and the text-based explanation of their evaluation. We study the empirical tradeoff between predictive and diagnostic abilities, in applying various methods to estimate this fundamental relationship. We incorporate methods previously employed in the marketing literature, and methods that are so far less common in the marketing literature. For generalizability, we analyze 25,241 products in nine product categories, and 260,489 reviews across five review platforms. We find that neural network-based machine learning methods, in particular pre-trained versions, offer the most accurate predictions, while topic models such as Latent Dirichlet Allocation offer deeper diagnostics. However, neural network models are not suited for diagnostic purposes and topic models are ill equipped for making predictions. Consequently, future selection of methods to process text reviews is likely to be based on analysts' goals of prediction versus diagnostics.

Published by Elsevier B.V.

1. Introduction

Over the past two decades, researchers, analysts, consumers and managers have witnessed an explosion of text data (Berger et al., 2020; Humphreys & Wang, 2018), what some call "big data" or "data deluge" (Bell, Hey, & Szalay, 2009; Borgman, 2015). It is estimated that 80–90% of today's data are unstructured (Harbert, 2021), and the ubiquitous text data constitute a great proportion of unstructured data. Part of this explosion of text data is attributed to the dramatic increase in consumer usage of e-commerce platforms (e.g., Amazon, eBay), consumer review sites (e.g., Yelp, TripAdvisor), and travel aggregators (e.g., Expedia, Booking.com), among others, whereby consumers have an option to provide a product (or service) review of their overall experience. Consumers can provide a product review through a numerical rating of their overall experience, which is commonly supplemented with a text-based explanation of their numerical rating. Such data have led to a rich body of marketing literature on consumer reviews.

Previous research in marketing has analyzed consumer review text using various natural language processing (NLP) methods, and we will conduct a review of these different methods to respond to Humphreys and Wang (2018)'s call to "discuss the methodological issues (or choices) consumer researchers face when dealing with text", as well as Berger et al.

https://doi.org/10.1016/j.ijresmar.2021.10.011
0167-8116/Published by Elsevier B.V.

(2020)'s call on researchers' capability to "extract underlying insight – to measure, track, understand, and interpret the causes and consequences of marketplace behavior". Some of these methods such as topic modelling have been widely employed in the marketing research literature, and others are supervised machine learning tools based on neural network models, recently developed in computer science and so far less commonly applied in marketing. By doing so, we intend to provide a large if not comprehensive assessment of tools for analyzing text reviews, on the tradeoff inherent in their use by marketing scholars and practitioners, which can guide method selection in future applications.

To assess the performance of different tools, we focus on the most fundamental relationship, i.e., the association between a consumer's numerical rating (sentiment) of their overall experience[1] as a dependent variable (DV), and the text-based explanation of their numerical rating as independent variables (IVs). Why do we focus on this fundamental relationship? For three reasons. First, the data are readily available for different contexts, which allow us to assess the generalizability of results across contexts. A review of the text-based marketing publications on consumer reviews provided in the next section indicates that several recently proposed methods in computer science can offer more accurate prediction of consumer sentiment but have so far been less common in the marketing literature. We would like to bring these methods to marketing scholars' attention by providing a comprehensive review of all methods in one single paper.

Second, with the readily available numerical rating associated with each text review, it is straightforward to evaluate the predictive ability of different methods on this same fundamental relationship we study. We acknowledge that the NLP tools we review can also be applied to analyze marketing questions and constructs beyond this fundamental relationship (e.g., studying concreteness, involvement, customer mindset metrics, etc.). However, many of these other constructs are context-dependent and specific and hence less comparable across contexts. For example, "involvement" is usually applicable for experience goods. In addition, it typically requires significant effort in coding these constructs of interest, which makes the comparison less generalizable. For example, Kübler et al. (2020) assess the performance of top-down language dictionaries and bottom-up machine learning approach (support vector machine) on predicting consumer mindset metrices from survey data, and find significant variations across contexts and product categories. Therefore, we consider the relationship we focus on to be an appropriate starting point for a comprehensive review of different tools, which would help researchers who are less familiar with NLP methods to select the appropriate tools for their research questions.

Third, the exploration of the relationship between the numerical rating and the text explanation provides essential diagnoses on why or why not a consumer is satisfied with the product or service. It affords new ways of reading the text (e.g., extracting features from text using NLP methods) and discoveries of systematic relationships in the text (e.g., the importance of a word, and associations between words) that scholars, analysts, managers or even consumers may overlook (Jurafsky, Ranganath, & McFarland, 2009). Such diagnoses can provide managerial insight on improving consumer satisfaction, for example, by prioritizing which areas to improve in order to enhance consumer experience.

For example, Chakraborty et al. (2021) employ machine learning to infer attribute ratings from the text, where diagnostics play a role in addition to prediction of the ratings. Likewise, we show for each of the nine product categories, how an analysis of text data provides inferences of topics related to consumer sentiment, as well as the words or lexical choices consumers employ to describe each aforementioned topic. The inferred topics and associated words are not completely unlike the diagnostics from models of consumer attitude, e.g., on beliefs and evaluative aspects (Fishbein & Ajzen, 1977), and conjoint and logit consumer preference and choice models, e.g., based on the presence, absence, and importance of attribute levels (e.g., Gensch & Recker, 1979; Green & Srinivasan, 1978). Consequently, sentiment models can provide diagnostics similar to attitude, conjoint or choice models, which can serve as useful inputs for enhancing product or service design, choice, usage, and post-use sentiment. We demonstrate in the Conclusion and Discussion sections how the determinant words and topics from sentiment models can quickly provide useful low-cost managerial inputs aimed at improving the design and marketing of products and services, and the corresponding customer experience.

To assess the performance of different tools in addressing this basic question, we define two evaluation metrics: predictive ability, and diagnostic ability. Predictive ability is defined based on the accuracy of the model's prediction of the consumer sentiment rating, measured by the F1-score which balances precision and recall, and the mean absolute and squared errors (MAE and MSE), on a test sample of reviews not employed to estimate or tune the model. Diagnostic ability is defined based on the managerial insights the model provides to improve the consumer experience, and is measured by the determinant words, i.e., words that are determinant or statistically significant in explaining the variance in positive and negative sentiment ratings.

Why is it important to study predictive and diagnostic abilities of methods? Predictive ability is important for stakeholders such as marketing scholars and practitioners to select from a large array of tools and models based on their quantitively measured accuracy (e.g., Abramson et al. 2000). Predictive ability is also useful for identifying customers who exhibit more desired behaviors (Andrews, Ainslie, & Currim, 2002, 2008) and sentiments than others. Diagnostic ability is important for practitioners and scholars who seek to determine the qualitative reasons for customer behavior (Andrews, Ansari, & Currim, 2002; Andrews, Currim, & Leeflang, 2011) or sentiment to improve the consumer experience. In sum, when there are many methods and tools available for analyzing text reviews, it is important to understand the quantitative and qualitative per-

[1] Although the numerical rating may not be equivalent to sentiment, in research using consumer review data, the numerical rating has been used widely as a proxy for consumer sentiment or review valence (e.g., Zhang and Godes, 2018).

formance of different tools with respect to different goals, predictive and diagnostic, to select the most appropriate model for analyzing the data in future applications.

While the natural language processing (NLP) literature in computer science has given sufficient attention on model comparison based on predictive ability, a review of this literature (more details in the next section) shows that diagnostic ability is an area that deserves more attention. Such a knowledge gap on the tradeoff can seriously limit methodological choices in future marketing applications conducted in scholarly and corporate settings (Berger et al., 2020), and there exist unmet needs regarding both the standard set of methods and the criteria for method selection (Humphreys and Wang, 2018).

The tradeoff between predictive and diagnostic abilities is not new to analyses of consumer reviews and has a rich forty-year history in market response modelling. A market response model, e.g., a consumer preference or conjoint (Andrews, Ansari, et al., 2002), consumer choice or multinomial logit (Andrews, Ainslie, et al., 2002, 2008), or sales response model (Andrews et al., 2011), which offers the best predictions may not offer the best diagnostics. In other words, it is likely there is no "winner takes all" when it comes to methods. This "no winner takes all" outcome can happen because while collinearity between independent variables (such as brands, prices, and promotions) can help with prediction, it may hurt diagnostic ability if the parameter estimates turn out to have signs opposite to expectation (Abramson et al., 2000). For example, the price coefficient which should be negative could turn out positive, and the promotion coefficient which should be positive could turn out negative. Such less-than-desirable diagnostic ability, even from a model which offers the best predictive ability, would signal that the model is not acceptable for managers or scholars, because the diagnostics it offers are opposite to expectations (Andrews, Ainslie, et al., 2002; Andrews, Ansari, et al., 2002; Andrews et al., 2008, 2011).

The "no winner takes all" outcome is similar to the "no free lunch" theorem in computer science, which suggests that there is no machine learning algorithm that dominates other algorithms in all contexts, and the choice of algorithm should be context dependent. While the "no free lunch" theorem focuses on predictive ability, we consider the diagnostic ability as an additional dimension of consideration in method selection. If sentiment analysis were only about classification on positivity or negativity, the star ratings would suffice. However, sentiment analysis goes beyond classification. As noted above, we are interested in the determinant words, empirically useful in explaining differences between higher (more positive) and lower (less positive) sentiment ratings, and inferred topics, labelled diagnostic ability. If positive words (adjectives) had negative coefficient signs and negative words had positive signs, even if the associated model were judged to have high predictive ability, the overall assessment of the model would be lower because of its lower diagnostic ability.

As we show later, some models offer a better understanding of the topics that the words are associated with, i.e., whether an adjective such as "good" in the hotel setting is associated with hotel or room amenities or the booking experience. And some models that offer good diagnostics are simply not suited for prediction. Consequently, it is important to evaluate sentiment models on both predictive and diagnostic abilities. Otherwise, we are left with the promise (or hype) of modern artificial intelligence (AI) methods without a clear knowledge of their inherent empirical benefits and drawbacks, relative to traditional methods.

2. Method

We follow two steps to select relevant methods for automated text-based sentiment analysis of consumer reviews. First, we conduct a review of published papers on text-reviews in the following journals (in alphabetical order) during 2011 to 2020: *Information Systems Research, International Journal of Research in Marketing, Journal of Consumer Research*, *Journal of Marketing, Journal of Marketing Research, Management Science,* and *Marketing Science.* This review provides an overview of research questions or relationships which have been addressed (the DVs and IVs) and findings, which are summarized in Online Appendix Table A1.

Some studies employ review text to extract review topics (e.g., Anderson & Simester, 2014; Kirmani et al., 2017; Puranam, Narayan, & Kadiyali, 2017; Wang & Chaudhry, 2018). For instance, Puranam et al. (2017) study changes in consumer opinion about health (and other topics) following the calorie posting regulation in New York City in 2008 by applying a topic model on restaurant reviews. They find the proportion of discussion on health topics increased after the regulation. Other studies employ review text to refine the valence of the review beyond the star rating (e.g., Ludwig et al., 2013; Tirunillai & Tellis, 2012, 2017; Wu, 2015). For instance, Tirunillai and Tellis (2012) obtain the valence of the reviews from Amazon.com, Epinions.com and Yahoo! Shopping, and assess the relationship between review valence and stock performance.

In addition, a few studies employ review text to identify the features of reviews (e.g., Ghose, Ipeirotis, & Li, 2012, 2019; Packard & Berger, 2017; Rubera, 2015; Zhang & Godes, 2018). For instance, Ghose et al. (2019) classify readability and subjectivity of text reviews, and extract hotel features from text reviews, which are both treated as factors that affect consumer utility. One fundamental question that gets little attention is whether review text can predict the numerical rating. This question requires us to assess the predictive ability across methods, determine the best model for prediction accuracy, and qualitatively diagnose or interpret the relationship, which is the focus of our paper.

In the second step, following Humphreys and Wang (2018), in Table 1, we categorize the methods employed in these papers into three broad categories: dictionary-based, classification, and topic discovery. We note that the main strength of dictionary-based methods is for behavioral scholars and researchers to conduct discipline-based theory testing, drawing theories developed in psychology, sociology, anthropology, and other social science disciplines (Humphreys & Wang, 2018). Behavioral researchers who are interested in specific theory-based constructs, variables, attributes, features, characteristics,

Table 1
A Summary of Text-based Sentiment Analysis Methods Selected for Comparison Purposes.

Method	Category of approach	Machine learning approach	Whether and who employs the method in marketing publications on review text
LIWC	Dictionary-based	Not machine learning	Chen, 2017; Goes, Lin, & Au Yeung, 2014; Hartmann et al., 2019; Ludwig et al., 2013; Ransbotham, Lurie, & Liu, 2019, Sridhar & Srinivasan, 2012; Van Laer et al., 2019; Villarroel Ordenes et al., 2017; Vermeer et al., 2019
SentiWordNet	Dictionary-based	Not machine learning	Archak, Ghose, & Ipeirotis, 2011; Ghose, Ipeirotis, & Li, 2019
K-Nearest Neighbors	Classification	Classic, Supervised, Non-Neural Network	Hartmann et al., 2019
Naïve Bayes	Classification	Classic, Supervised, Non-Neural Network	Ghose, Ipeirotis, & Li, 2012; Hartmann et al., 2019; Tirunillai & Tellis, 2012, 2017; Vermeer et al., 2019; Liu, Lee, & Srinivasan, 2019
SVM	Classification	Classic, Supervised, Non-Neural Network	Hartmann et al., 2019; Tirunillai & Tellis, 2012, 2017; Vermeer et al., 2019; Liu, Lee, & Srinivasan, 2019; Bai et al., 2020
Logistic Regression	Classification	Classic, Supervised, Non-Neural Network	Vermeer et al., 2019
Ordered Logistic Regression	Classification	Classic, Supervised, Non-Neural Network	No
Decision Tree	Classification	Classic, Supervised, Non-Neural Network	No
Random Forest	Classification	Classic, Supervised, Non-Neural Network	Hartmann et al., 2019
XGBoost	Classification	Classic, Supervised, Non-Neural Network	No
AdaBoost	Classification	Classic, Supervised, Non-Neural Network	No
Feedforward NN	Classification	Neural Network Supervised Learning	Hartmann et al., 2019
LSTM	Classification	Neural Network Supervised Learning	Liu, Lee, & Srinivasan, 2019
BiLSTM	Classification	Neural Network Supervised Learning	No
CNN	Classification	Neural Network Supervised Learning	Liu, Lee, & Srinivasan, 2019; Timoshenko & Hauser, 2019
CNN-LSTM	Classification	Neural Network Supervised Learning	No
FastText	Classification	Neural Network Supervised Learning	No
BERT	Classification	Neural Network Supervised Learning Pre-trained	Bai et al., 2020
RoBERTa	Classification	Neural Network Supervised Learning Pre-Trained	No
ALBERT	Classification	Neural Network Supervised Learning Pre-Trained	No
DistilBERT	Classification	Neural Network Supervised Learning Pre-Trained	No
XLNet	Classification	Neural Network Supervised Learning Pre-Trained	No
LDA	Topic discovery	Unsupervised, Non-Neural Network	Büschken & Allenby, 2016; Lee & Bradlow, 2011; Puranam, Narayan, & Kadiyali, 2017; Tirunillai & Tellis, 2014; Wang & Chaudhry, 2018

or phenomena (e.g., emotional arousal and valence, and sentiment expression) can identify and scale the corresponding words from pre-defined dictionaries available in commercial software (e.g., Linguistic Inquiry and Word Count (LIWC), Revised Dictionary of Affect in Language (RDAL), SentiWordNet, Stanford Sparser). Alternatively, behavioral researchers can identify and scale attributes that are of particular interest but not present in a pre-defined dictionary (e.g., competence, warmth, design innovativeness) based on manual coding (e.g., Kirmani et al., 2017; Rubera, 2015).

In summary, dictionary-based methods are suited for theory-testing and scaling related to any social science-based theory, classification procedures are suited for prediction, and topic discovery methods are suited for managerial decision-oriented diagnostics which are not necessarily theory-based. Because dictionary-based methods are rarely if ever employed for predictive purposes, and topic discovery methods are not capable of prediction, our model comparison on predictive ability only focuses on classification procedures.

Different from dictionary-based methods, classification and topic discovery-based approaches do not begin with predefined words in a dictionary. Classification employs supervised learning methods to classify the dependent variable, including the overall numerical rating or review characteristics, such as subjectivity (Ghose et al., 2012, 2019), informativeness (Zhang & Godes, 2018), consumer endorsement style (Packard & Berger, 2017) and valence (Tirunillai & Tellis, 2012, 2017). Researchers have applied classic non-neural network-based methods such as Naïve Bayes, Support Vector Machines (SVM) and other neural network-based learning models such as Convolutional Neural Network (CNN), employing open-source Python or R packages or commercial packages (such as IBM's Alchemy API).

We choose classic non-neural network-based methods that have been applied in the marketing publications (Naïve Bayes, SVM, Logistic Regression and Random Forest), as well as several others that have not been applied in the marketing publications on consumer reviews (Ordered Logistic Regression, Decision Tree, XGBoost and AdaBoost).[2] In addition, we consider the few modern neural network-based learning methods that have been applied in marketing publications (Feedforward Neural Network, Convolutional Neural Network (CNN), and Long Short Term Memory (LSTM)), and complement these methods with several other recently proposed methods in computer science (BiLSTM, CNN-LSTM, and FastText). Importantly, we include recently developed pre-trained methods (BERT, RoBERTa, ALBERT, DistilBERT, and XLNet) that are suitable for text analysis. Although some of these methods are known to marketing researchers, to our best knowledge, only one paper (Bai et al., 2020) published in the journals we surveyed has applied BERT, and we would like to highlight the performance of these methods for marketing scholars. The inclusion of the non-neural network machine learning methods, like Chakraborty et al. (2021), allows a determination of the incremental value of neural network-based methods over the classic non-neural network-based machine learning methods.

Finally, the topic discovery approach relies on unsupervised (versus supervised as in all classification approaches) learning methods to extract topics from the reviews. The Latent Dirichlet Allocation (LDA) is the most applied topic discovery approach in marketing applications (e.g., Puranam et al., 2017; Tirunillai & Tellis, 2014; Wang & Chaudhry, 2018). We include LDA for diagnostic purposes, acknowledging that it is not suitable for prediction.

We retain the three categories of methods proposed by Humphreys and Wang (2018). Two of the three categories, the dictionary-based method and topic discovery, do not have a dependent variable, while the classification method does. We choose not to combine the dictionary-based method and topic discovery method into one broader category because they are fundamentally different: topic discovery is based on machine learning and can, a posteriori, be improved with data to achieve practical insights about consumer sentiment, while the dictionary-based method is largely dependent on researchers' theory development and testing motivations, whereby researchers a priori select and scale the words from predetermined dictionaries.

Table 1 summarizes the methods for sentiment analysis, including those selected for comparison, the category of approaches that the methods are based on (dictionary, classification, or topic discovery), the category of the method within machine learning approaches (neural network-based or not, pre-trained or not), and whether the method has, to our best knowledge, been employed in marketing publications on text reviews. For methods employed in marketing publications on text reviews, we also list papers that apply the corresponding method.[3] Details on each method are described in Online Appendix A.

Among the papers summarized in Table 1, only four papers make comparisons between their focal approach and at least one other approach, and another two papers conduct reviews on multiple methods. Timoshenko and Hauser (2019) apply CNN to identify customer needs from consumer reviews and compare this approach with a traditional manual processing approach. Liu, Lee, and Srinivasan (2019) compare performance of conventioanl classifiers including SVM and Naïve Bayes to the performance of neural network models such as the recurrent neural networks LSTM, the recursive neural networks, and CNN. Archak, Ghose, and Ipeirotis (2011) compare their proposed approach to extract most popular opinions from consumer reviews with a manual processing approach. Büschken and Allenby (2016) propose a new LDA model that utilizes the sentence structure in the reviews and demonstrate its superiority relative to existing LDA models in inference and prediction of consumer ratings.

In a review paper, Hartmann et al. (2019) compare LIWC (and Vader), SVM, Naïve Bayes, Feedforward NN and LDA. However, like previous studies, they do not consider many of the neural network methods in our comparison, many of which provide the best predictions, perhaps because they focus on methods which have been applied in the marketing literature. Likewise, Vermeer et al. (2019) compare LIWC (and other dictionary-based methods and manual coding), SVM, (Multinomial and Bernouli) Naïve Bayes, Logistic Regression, Gradient Descent and Passive Aggressive, but they too do not consider any of the neural network methods. Different from these two review papers, Heitmann et al. (2020) take into consideration neural network models in a large-scale meta-analysis on the sentiment analysis accuracy across methods, including neural network, Naïve Bayes, SVM, Logistic regression, random forest, and lexicon. However, they focus on predictive ability and do not differentiate between different variants of neural network models. Consequently, we complement existing studies by exploring the tradeoff (prediction and diagnosis) between a large set of systematically (carefully) chosen NLP methods in the same context, i.e., the same model and data setting.

[2] Note that although these methods have not been applied in marketing publications on consumer reviews, some of them have been applied in marketing publications on other topics such as advertising (Rafieian & Yoganarasimhan, 2021) and search personalization (Yoganarasimhan, 2020).

[3] A paper can be listed under multiple methods if more than one method is applied in the paper.

While Table 1 (col. 4) focuses on methods employed in the marketing publications, studies in computer science have largely focused on proposing methods and assessing their predictive performance on consumer sentiment ratings (either binary positive/negative sentiment, or fine-grained star rating) on datasets from Amazon, IMDB, Yelp, etc. Online Appendix Table A2 reviews a set of computer science studies that compare methods on predicting consumer ratings with Amazon and Yelp datasets. Each of these studies introduces or focuses on a certain method (e.g., BERT, CNN, FastText) and compares the predictive ability of the chosen method with several variants of the method or a few other methods in a related category (e.g., neural network models). None of these studies to our best knowledge comes remotely close to conducting the scope of our investigation of models both within and across categories as shown in Table 1. In addition, comparisons on diagnostic ability, i.e., the words that empirically discriminate between better and worse consumer sentiment, and topics inferred, have traditionally received much less attention from computer science researchers, although diagnostics are now getting more attention (e.g., Google's newly developed Language Interpretability Tool (LIT)[4]).

We select multiple datasets on consumer reviews that contain both text reviews and numerical ratings. For each dataset, we convert text data into two dimensional arrays of quantitative values,[5] and apply each of the classification methods in Table 1[6] to obtain relationships between review text and numerical ratings. Subsequently, we evaluate the predictive and diagnostic abilities of different methods.

For predictive ability, we use the review text to predict the original numerical review rating using each of the classification methods. In all analyses, we separate the data into 3 sets: (1) training, (2) tuning, and (3) test. The training set comprises 2/3 of the data, while the tuning and test sets are 1/6 each for a total of 1/3 of the data (Steckel & Vanhonacker, 1993). The training set is the data on which the model is trained/estimated. The tuning set is used for tuning hyper-parameters. Once the model is estimated and tuned, we test its performance on the test set as is standard in the literature.

For diagnostic ability, five of the twenty-three methods directly enable diagnostics: LIWC, SentiWordNet, Logistic Regression, Ordered Logistic Regression and Topic Discovery. The diagnostics from LIWC and SentiWordNet are based on a predetermined dictionary. Logistic Regression and Ordered Logistic Regression explore dependence between words and overall numerical ratings, and consequently can identify the determinant words which are most predictive of overall numerical ratings. In contrast, the topic discovery method identifies topics based on an analysis of interdependence between words. Although the neural network models are structurally complex with multiple layers of variables for deep learning, and thus do not afford diagnostic abilities, recent work on Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) can be used to obtain basic diagnoses for neural network models (Rai, 2020), such as what are the most predictive words for a certain outcome (i.e., sentiment rating).

Our expectations on predictive and diagnostic abilities are as follows. Regarding predictive ability, we expect neural network-based machine learning methods to provide better predictions than non-neural network-based machine learning methods. While all machine learning methods are able to learn complex patterns in the available text data, with sufficient data and effort in training, neural network models are expected to learn the complex patterns in a particular context better because of their more complex architecture, i.e., hidden layers which promote deep learning, (Bai et al., 2020; Hartmann et al. 2019; Liu, Lee & Srinivasan, 2019; Timoshenko & Hauser, 2019).

Regarding diagnostic ability, we expect that classification methods (and especially neural network methods), which "sacrifice transparency and interpretability for prediction accuracy" (Rai, 2020), may not offer diagnostics at the same level as topic discovery methods, which are designed for diagnostics instead of prediction (Büschken & Allenby, 2016; Lee & Bradlow, 2011; Puranam, Narayan, & Kadiyali, 2017; Tirunillai & Tellis, 2014; Wang & Chaudhry, 2018). Take hotel reviews as an example, if classification approaches identify a set of determinant positive adjectives or words such as "great", "wonderful", "excellent", and "perfect", it becomes difficult to judge whether the consumer, when employing such words, is referring to hotel or room amenities, service, the booking experience, or the location.

In contrast to classification approaches, topic discovery methods account for interdependence between the words (Büschken & Allenby, 2016; Wang & Chaudhry, 2018). The explicit consideration of interdependence between words can lead to the identification of topics. In other words, it is possible to infer that the word "clean" is associated with other words such as "breakfast", "parking", "pool", "area", and "free", which are related to the topic of hotel amenities. And that the word "great" is associated with another set of words such as "thank", "enjoyed", "staff", "review", "time", and "happy", which are associated with the topic of positive experience. Further, such identification of topics can provide better guidance for

[4] We thank the review team for suggesting this tool.

[5] We convert text data to quantitative values with the following packages: For KNN, Naïve Bayes, SVM, Logistic Regression, Ordered Logistic Regression, Decision Tree, Random Forest, and Feedforward Neural Net, we use TfidfVectorizer from scikit-learn. For LDA, we use CountVectorizer from scikit-learn. For Feedforward NN, CNN, LSTM, BiLSTM, CNN-LSTM, and FastText, we use Keras to convert texts to sequences of integers with equal lengths. We employ the uncased versions for BERT, RoBERTa, ALBERT, and the cased version for XLNet from huggingface's transformers package via simpletransformers, which converts text to numerical data.

[6] We employ multiple Python libraries to build our models. Specifically, we use (i) Keras including its code examples for Feedforward Neural Network, LSTM, BiLSTM, CNN, CNN-LSTM, and FastText, (ii) scikit-learn for LDA, Naïve Bayes, K-Nearest Neighbors (KNN), SVM, Random Forest, Logistic Regression, Decision Tree, and AdaBoost, (iii) mord for ordered logistic regression, (iv) XGBoost for XGBoost, and (v) simpletransformers (a wrapper for the transformers library from huggingface) for BERT, RoBERTA, DistilBERT, ALBERT, and XLNet. For KNN, we acknowledge the possibility of ties, which will affect the results. For Naïve Bayes, we employ the multinomial distribution. For SVM, the input matrix is weighted by term-frequency inverse-document frequency (TF-IDF), which gives higher weights to uncommon words across reviews rather than treating all the words as equal, a common practice for text analysis tasks.

analysts, managers and scholars, towards understanding and improving the consumer experience with products and services.[7]

3. Data

To ensure generalizability, we collect datasets on nine product categories (hotels, airlines, drugs, books, automotive, office products, patio, musical instruments, and instant video) from five review sites in two different settings. The first setting is review sites, where consumers can provide both ratings and text-based explanations of their ratings. The corresponding eight product categories are from Tripadvisor.com[8] (provided by Datafiniti's Business Database), drugs.com[9] (Gräßer et al., 2018), Goodreads[10] (Wan & McAuley, 2018; Wan et al., 2019), and Amazon[11] (He & McAuley, 2016; McAuley, Targett, Shi, & van den Hengel, 2015). The second setting is social media, where consumers provide sentiment information in words without a numerical rating. The corresponding product category is airlines from Twitter[12] (the original source is Crowdflower's Data for Everyone library).

The two settings vary in three aspects: i) Ratings are provided on review sites but not on Twitter; ii) Twitter has a limitation on the number of words whereas there is no such limit in the other setting; and iii) in the Twitter setting, consumers are communicating directly with airlines via mentions, while in the other eight datasets the communications are not aimed directly at firms. Because consumers do not provide numerical ratings on Twitter, the numerical ratings associated with Tweets are coded by other researchers, and reflect the sentiment (negative, neutral, or positive). Therefore, the data-generating process of the Twitter data differs from that of the other datasets, which allows us to enhance the generalizability of our results.

Table 2 provides summary statistics on the nine datasets. The numerical sentiment rating is on a 3-point scale for the airline category, on a 10-point scale for the drug category, and on a 5-point scale for the other seven product categories. The number of products per category ranges from 6 (airlines) to 15,000 (books). The number of reviews per dataset ranges from 10,000 (hotels) to 62,348 (drugs). The number of unique words in the text-based information ranges from 14,934 (airlines) to 61,106 (books). The average number of reviews per product ranges from 2.6 (books) to 22 (office products, instant videos), and 2,440 (airlines). The average number of words per review ranges from 10.5 (airlines) to 44.1 (automotive products) and 79.7 (patio products).

The standard deviations of the ratings, the number of reviews per item, and the number of words per review are substantial relative to the corresponding means. Consequently, we have a good variation across the nine product categories on the number of products, reviews, unique words, average number of reviews and average number of words per review for our findings to be generalizable. Therefore, these datasets are suited to address our research questions.[13]

We only consider datasets which have both sentiment ratings as well as text-based explanations because the focus of this work is to compare the predictive and diagnostic abilities across methods based on the fundamental relationship between review text and sentiment ratings. We do not consider research questions outside of this domain that can be addressed using these datasets because of the large scope (number of models and datasets) of our investigation. Future work can focus on fewer models calibrated on different phenomena (e.g., social movements) in settings that contain qualitative text data.

4. Results

We implement the methods discussed in Section 2 on each of the nine datasets discussed in Section 3. For neural network-based machine learning models, we select the number of epochs that leads to the best prediction. Below we discuss related results on predictive ability in Section 4.1 and on diagnostic ability in Section 4.2.

4.1. Predictive ability

We assess predictive ability with five measures -- F1, Precision, Recall, Mean Absolute Error (MAE), and Mean Squared Error (MSE) -- all on a test sample which is not used for model estimation and hyper-parameter tuning.

F1, Precision, and Recall are defined as follows. In binary classification (when the outcome is 1 or 0), precision is the proportion of predicted 1's that are truly 1, and recall is the proportion of actual 1's that is correctly classified as "1". F1 is a weighted harmonic mean of precision and recall. Our application is a multi-class classification because there are 3, 5 or

[7] The interested reader is referred to "aspect-based sentiment analysis" in computer science, which links sentiment to specific product/service characteristics (https://paperswithcode.com/task/aspect-based-sentiment-analysis).

[8] https://www.kaggle.com/datafiniti/hotel-reviews

[9] https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018

[10] https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/reviews

[11] https://nijianmo.github.io/amazon/index.html

[12] https://www.kaggle.com/crowdflower/twitter-airline-sentiment

[13] Note that because the machine learning models are trained on context-specific data, the relationship between text and rating is context dependent. However, we aim to generalize the relative performance of different methods by applying them in different contexts (i.e., product categories).

Table 2
Summary Statistics for Nine Datasets.

Product Category (Platform)	Hotel (Tripadvisor)	Airline (Twitter)	Drug (drugs.com)	Books (Goodreads)	Automotive (Amazon)	Office (Amazon)	Patio (Amazon)	Music Instrument (Amazon)	Instant Video (Amazon)
Total number of products (e.g., hotels, airlines, etc.)	1,433	6	1,000	15,000	1,835	2,420	962	900	1,685
Total number of reviews	10,000	14,640	62,348	39,160	20,467	53,237	13,258	10,254	37,125
Number of unique words	21,999	14,934	34,603	61,106	27,665	53,256	28,977	20,436	53,903
Range of rating	1–5	1–3	1–10	1–5	1–5	1–5	1–5	1–5	1–5
Mean of rating	4.1	1.5	7	4.1	4.5	4.3	4.2	4.5	4.2
Standard deviation of rating	1.2	0.8	3.3	1.1	0.9	0.9	1.1	0.9	1.1
Frequency of rating 10	-	-	19,312 (31.0%)	-	-	-	-	-	-
Frequency of rating 9	-	-	10,607 (17.0%)	-	-	-	-	-	-
Frequency of rating 8	-	-	7,331 (11.8%)	-	-	-	-	-	-
Frequency of rating 7	-	-	3,725 (6.0%)	-	-	-	-	-	-
Frequency of rating 6	-	-	2,514 (4.0%)	-	-	-	-	-	-
Frequency of rating 5	4,840 (48.4%)	-	3,245 (5.2%)	18,689 (47.7%)	13,926 (68.0%)	30,318 (56.9%)	7,029 (53.0%)	6,932 (67.6%)	20,890 (56.3%)
Frequency of rating 4	2,849 (28.5%)	-	1,978 (3.2%)	11,224 (28.7%)	3,964 (19.4%)	15,008 (28.2%)	3,382 (25.5%)	2,083 (20.3%)	8,446 (22.8%)
Frequency of rating 3	1,190 (11.9%)	2,363 (16.1%)	2,628 (4.2%)	5,520 (14.1%)	1,430 (7.0%)	5,058 (9.5%)	1,657 (12.5%)	772 (7.5%)	4,186 (11.3%)
Frequency of rating 2	554 (5.5%)	3,099 (21.2%)	2,596 (4.2%)	2,332 (6.0%)	605 (3.0%)	1,724 (3.2%)	673 (5.1%)	250 (2.4%)	1,885 (5.1%)
Frequency of rating 1	567 (5.7%)	9,178 (62.7%)	8,412 (13.5%)	1,395 (3.6%)	542 (2.6%)	1,129 (2.1%)	517 (3.9%)	217 (2.1%)	1,718 (4.6%)
Average number of reviews per product	7	2,440	62.3	2.6	11.2	22	13.8	11.4	22
Standard deviation of number of reviews per product	18.4	1,002.6	261.2	13.7	11.2	27.5	16.8	12.9	38.7
Average number of words per review	61.2	10.5	45	54.5	44.1	75.3	79.7	45.9	48.1
Standard deviation of number of words per review	55.3	4.1	23.4	78.3	52.8	87.6	77.9	58.2	78.3

10 outcomes (classes) or numerical rating scales across product categories. Precision, Recall and F1 are applied to each rating independently, and the reported numbers are the weighted average of the measures (across ratings) on each outcome.

Table 3 Panel A presents the performance of all twenty classification methods in predicting the original review rating (1–3 for airlines, 1–10 for drugs, and 1–5 for the other seven product categories), across nine product categories, sorted on F1 because it is the metric most commonly employed in prediction. The last row, "Frequent Class", refers to a baseline method which predicts the modal or most frequent rating for all reviews. To check for convergence or divergence in the model per-

Table 3
Accuracy of Model Predictions of Sentiment Rating on Test Samples.

Panel A. Mean and Standard Deviation of Predictive Accuracy Across the Nine Product Category Datasets (sorted on F1, shaded cells indicate pre-trained methods that appear on top 5)

Model	F1	Precision	Recall	MAE	MSE
BERT	61.03 (8.71)	60.11 (8.55)	63.25 (9.37)	0.5526 (0.3010)	1.2717 (1.5473)
XLNet	60.73 (8.59)	59.88 (8.38)	63.07 (9.18)	0.5572 (0.2975)	1.2906 (1.5259)
DistilBERT	60.08 (8.71)	59.05 (8.51)	62.44 (9.34)	0.5734 (0.3162)	1.3513 (1.6636)
RoBERTa	59.47 (9.24)	58.10 (9.45)	62.78 (9.63)	0.5663 (0.2918)	1.3051 (1.4647)
SVM	58.63 (9.51)	58.63 (9.58)	62.99 (9.21)	0.6202 (0.4107)	1.7338 (2.5665)
FastText	58.16 (10.89)	57.24 (10.26)	62.81 (9.73)	0.6277 (0.4371)	1.7860 (2.7433)
ALBERT	58.05 (9.19)	56.82 (9.35)	61.80 (9.78)	0.5907 (0.3070)	1.3828 (1.5767)
CNN-LSTM	57.94 (11.09)	57.29 (11.16)	61.49 (9.66)	0.6134 (0.3968)	1.6053 (2.3634)
Logistic Regression	57.56 (10.41)	57.70 (10.68)	62.82 (9.55)	0.6408 (0.4420)	1.8609 (2.8117)
BiLSTM	57.30 (10.72)	56.87 (10.36)	61.16 (10.12)	0.6194 (0.3951)	1.6127 (2.3284)
XGBoost	57.23 (7.97)	58.32 (7.55)	62.26 (8.09)	0.6507 (0.3981)	1.8791 (2.7013)
CNN	57.20 (10.79)	55.76 (11.06)	60.89 (10.05)	0.6413 (0.4186)	1.7199 (2.4771)
Ordered Logistic Regression	56.93 (10.81)	57.33 (10.25)	58.41 (11.95)	0.6164 (0.4457)	1.4182 (2.1958)
Feedforward NN	56.43 (9.43)	54.11 (9.66)	61.49 (9.41)	0.6315 (0.4023)	1.6956 (2.4791)
LSTM	56.18 (10.29)	54.59 (11.12)	61.09 (9.69)	0.6369 (0.4198)	1.7409 (2.6401)
AdaBoost	52.70 (11.33)	51.72 (11.54)	59.03 (10.70)	0.7524 (0.5731)	2.4047 (3.9546)
Multinominal Naïve Bayes	52.57 (9.93)	53.69 (9.38)	59.80 (9.98)	0.7373 (0.5322)	2.3405 (3.7366)
Decision Tree	50.19 (8.92)	49.64 (9.10)	57.44 (10.16)	0.8053 (0.5592)	2.5442 (3.8232)
Random Forest	49.18 (9.43)	60.03 (8.30)	59.48 (8.73)	0.7839 (0.5038)	2.5550 (3.7430)
KNN	48.99 (11.63)	48.74 (11.36)	57.85 (11.02)	0.7973 (0.5905)	2.5985 (4.1649)
Frequent Class	39.80 (12.76)	31.51 (11.75)	55.08 (11.49)	0.9539 (0.7802)	3.5309 (6.0528)

Panel B: F1-based Predictive Accuracy Results for Each Product Category Ordered from Highest to Lowest Accurate Method (shaded cells indicate pre-trained methods that appear on top 5)

Order	Hotel	Airline	Drug	Book	Automotive	Office	Patio	Music instrument	Instant video
1	XLNet (62.29%)	RoBERTa (79.17%)	BERT (45.75%)	BERT (58.67%)	BERT (65.62%)	FastText (61.10%)	XLNet (57.62%)	XLNet (61.79%)	BERT (61.53%)
2	BERT (61.69%)	BERT (78.99%)	ALBERT (45.30%)	DistilBERT (57.08%)	XLNet (64.96%)	CNN-LSTM (60.21%)	SVM (57.39%)	SVM (61.20%)	DistilBERT (60.59%)
3	RoBERTa (61.32%)	CNN-LSTM (78.56%)	XLNet (45.29%)	CNN (56.93%)	DistilBERT (64.27%)	XLNet (60.00%)	BERT (56.38%)	Ordered Logit (61.00%)	RoBERTa (60.58%)
4	ALBERT (61.25%)	ALBERT (78.55%)	RoBERTa (44.74%)	XLNet (56.52%)	BiLSTM (63.98%)	BERT (59.76%)	DistilBERT (56.17%)	DistilBERT (60.94%)	XLNet (60.01%)
5	DistilBERT (61.22%)	XLNet (78.10%)	XGBoost (44.72%)	CNN-LSTM (55.69%)	RoBERTa (63.72%)	XGBoost (59.27%)	FastText (55.44%)	BERT (60.82%)	FastText (59.20%)
6	SVM (58.75%)	BiLSTM (78.02%)	DistilBERT (44.43%)	RoBERTa (54.61%)	Ordered Logit (63.33%)	RoBERTa (59.11%)	Ordered Logit (55.40%)	CNN-LSTM (60.41%)	CNN (58.07%)
7	FastText (58.43%)	FastText (77.81%)	FNN (41.21%)	Logit (54.60%)	CNN-LSTM (63.24%)	Logit (58.40%)	Logit (55.02%)	XGBoost (60.22%)	Logit (58.04%)
8	Logit (58.21%)	DistilBERT (77.81%)	SVM (40.55%)	SVM (54.20%)	CNN (63.13%)	Ordered Logit (58.37%)	ALBERT (54.98%)	BiLSTM (60.17%)	Ordered Logit (58.02%)
9	CNN-LSTM (57.89%)	SVM (77.30%)	RF (39.58%)	LSTM (53.54%)	SVM (62.88%)	DistilBERT (58.28%)	BiLSTM (54.57%)	FastText (60.08%)	CNN-LSTM (57.76%)
10	Ordered Logit (56.98%)	CNN (76.99%)	DT (37.76%)	Ordered Logit (53.24%)	FastText (62.51%)	BiLSTM (58.03%)	RoBERTa (54.02%)	ALBERT (60.02%)	SVM (57.42%)
11	BiLSTM (56.78%)	Logit (76.77%)	BiLSTM (37.37%)	FastText (52.97%)	Logit (62.20%)	SVM (57.99%)	CNN (53.65%)	CNN (59.66%)	ALBERT (56.27%)
12	CNN (55.32%)	FNN (76.75%)	LSTM (36.93%)	BiLSTM (52.81%)	XGBoost (62.08%)	FNN (57.47%)	XGBoost (53.58%)	Logit (58.65%)	LSTM (55.53%)
13	LSTM (55.14%)	LSTM (76.70%)	CNN-LSTM (36.22%)	XGBoost (51.88%)	FNN (61.44%)	AdaBoost (56.62%)	FNN (53.53%)	LSTM (58.44%)	XGBoost (55.28%)
14	MNB (55.06%)	XGBoost (73.41%)	Logit (36.15%)	FNN (51.70%)	LSTM (60.99%)	LSTM (56.55%)	LSTM (51.77%)	RoBERTa (57.99%)	AdaBoost (54.94%)
15	XGBoost (54.58%)	Ordered Logit (73.36%)	FastText (35.87%)	AdaBoost (51.10%)	ALBERT (59.59%)	ALBERT (56.21%)	CNN-LSTM (51.50%)	FNN (57.24%)	BiLSTM (53.98%)
16	FNN (54.57%)	MNB (72.51%)	CNN (35.29%)	ALBERT (50.32%)	DT (59.04%)	CNN (55.76%)	MNB (50.13%)	DT (56.72%)	FNN (53.91%)
17	AdaBoost (51.08%)	AdaBoost (71.46%)	MNB (34.95%)	MNB (47.28%)	AdaBoost (58.88%)	DT (52.62%)	KNN (46.68%)	AdaBoost (56.00%)	MNB (50.52%)
18	KNN (48.32%)	KNN (70.58%)	Ordered Logit (32.67%)	DT (44.49%)	MNB (57.55%)	MNB (50.06%)	AdaBoost (44.69%)	RF (55.19%)	KNN (46.93%)
19	RF (45.84%)	RF (69.08%)	AdaBoost (29.55%)	KNN (42.21%)	RF (55.83%)	RF (49.37%)	DT (43.95%)	KNN (55.12%)	DT (46.70%)
20	DT (44.72%)	DT (65.72%)	KNN (27.12%)	RF (41.52%)	KNN (55.57%)	KNN (48.37%)	RF (43.01%)	Frequent (55.04%)	RF (43.25%)
21	Frequent (31.19%)	Frequent (48.03%)	Frequent (15.05%)	Frequent (31.25%)	Frequent (55.12%)	Frequent (42.40%)	Frequent (39.38%)	MNB (55.04%)	Frequent (40.77%)

Panel C: MAE-based Predictive Accuracy Results for Each Product Category Ordered from Highest to Lowest Accurate Method (shaded cells indicate pre-trained methods that appear on top 5)

Order	Hotel	Airline	Drug	Book	Automotive	Office	Patio	Music Instrument	Instant Video
1	XLNet (0.4217)	BERT (0.2537)	RoBERTa (1.3020)	BERT (0.5108)	BERT (0.4475)	FastText (0.4651)	Ordered Logit (0.5425)	Ordered Logit (0.4576)	BERT (0.4879)
2	RoBERTa (0.4301)	RoBERTa (0.2545)	XLNet (1.3189)	XLNet (0.5387)	XLNet (0.4552)	Ordered Logit (0.4730)	XLNet (0.5507)	BERT (0.4640)	DistilBERT (0.4969)
3	ALBERT (0.4367)	ALBERT (0.2553)	BERT (1.3227)	DistilBERT (0.5430)	Ordered Logit (0.4555)	CNN-LSTM (0.4741)	SVM (0.5597)	DistilBERT (0.4675)	RoBERTa (0.4997)
4	BERT (0.4403)	CNN-LSTM (0.2631)	ALBERT (1.3596)	CNN-LSTM (0.5545)	BiLSTM (0.4657)	XLNet (0.4799)	BERT (0.5652)	XLNet (0.4705)	Ordered Logit (0.5090)
5	DistilBERT (0.4421)	XLNet (0.2689)	DistilBERT (1.3855)	Ordered Logit (0.5569)	RoBERTa (0.4681)	BERT (0.4813)	BiLSTM (0.5738)	SVM (0.4728)	XLNet (0.5102)
6	Ordered Logit (0.4763)	FastText (0.2697)	CNN-LSTM (1.6413)	CNN (0.5580)	DistilBERT (0.4710)	BiLSTM (0.4859)	DistilBERT (0.5756)	CNN-LSTM (0.4769)	FastText (0.5223)
7	SVM (0.4823)	SVM (0.2713)	BiLSTM (1.6436)	RoBERTa (0.5660)	SVM (0.4766)	FNN (0.4946)	FastText (0.5787)	ALBERT (0.4786)	CNN-LSTM (0.5368)
8	CNN-LSTM (0.4835)	BiLSTM (0.2734)	FNN (1.6746)	Logit (0.5669)	CNN (0.4789)	RoBERTa (0.4967)	RoBERTa (0.5846)	FastText (0.4786)	LSTM (0.5465)
9	FastText (0.4883)	DistilBERT (0.2738)	XGBoost (1.6828)	LSTM (0.5670)	FNN (0.4801)	XGBoost (0.4981)	Logit (0.5937)	XGBoost (0.4851)	Logit (0.5498)
10	BiLSTM (0.4973)	FNN (0.2770)	SVM (1.6889)	BiLSTM (0.5719)	Logit (0.4839)	Logit (0.4987)	FNN (0.6009)	BiLSTM (0.4862)	ALBERT (0.5556)
11	Logit (0.5045)	Logit (0.2795)	LSTM (1.7294)	SVM (0.5744)	FastText (0.4839)	SVM (0.4997)	CNN (0.6023)	LSTM (0.4927)	SVM (0.5558)
12	FNN (0.5183)	LSTM (0.2811)	CNN (1.7308)	FNN (0.5774)	LSTM (0.4851)	LSTM (0.5024)	CNN-LSTM (0.6041)	RoBERTa (0.4950)	FNN (0.5630)
13	LSTM (0.5207)	CNN (0.2848)	FastText (1.7665)	FastText (0.5960)	CNN-LSTM (0.4862)	DistilBERT (0.5049)	LSTM (0.6068)	Logit (0.4956)	CNN (0.5737)
14	CNN (0.5231)	Ordered Logit (0.2898)	Ordered Logit (1.7872)	ALBERT (0.5981)	XGBoost (0.4909)	ALBERT (0.5212)	ALBERT (0.6072)	CNN (0.4968)	BiLSTM (0.5766)
15	MNB (0.5321)	XGBoost (0.3176)	Logit (1.7948)	XGBoost (0.6199)	ALBERT (0.5035)	CNN (0.5235)	XGBoost (0.6213)	FNN (0.4980)	XGBoost (0.5882)
16	XGBoost (0.5525)	MNB (0.3254)	RF (2.0764)	AdaBoost (0.6449)	DT (0.5229)	AdaBoost (0.5385)	MNB (0.6670)	AdaBoost (0.5120)	AdaBoost (0.6107)
17	AdaBoost (0.6125)	KNN (0.3406)	MNB (2.1260)	MNB (0.6832)	AdaBoost (0.5237)	DT (0.5832)	KNN (0.7118)	DT (0.5184)	MNB (0.6666)
18	KNN (0.6461)	AdaBoost (0.3418)	AdaBoost (2.2531)	KNN (0.7516)	MNB (0.5270)	MNB (0.5868)	AdaBoost (0.7348)	RF (0.5214)	KNN (0.7233)
19	RF (0.6989)	RF (0.3570)	DT (2.2573)	DT (0.7593)	RF (0.5343)	RF (0.5916)	DT (0.7353)	KNN (0.5219)	DT (0.7338)
20	DT (0.7385)	DT (0.3988)	KNN (2.3347)	RF (0.7708)	KNN (0.5363)	KNN (0.6093)	RF (0.7380)	Frequent (0.5219)	RF (0.7665)
21	Frequent (0.9034)	Frequent (0.5365)	Frequent (2.9962)	Frequent (0.8862)	Frequent (0.5402)	Frequent (0.6387)	Frequent (0.7674)	MNB (0.5219)	Frequent (0.7946)

formance across product categories, Panels B and C present the performance of all methods, per product category, sorted respectively on F1 and MAE. The results of the other fit statistics for each product category are available from the authors.

Three main results on predictive accuracy are as follows. First, Table 3 Panel A shows that several neural network-based learning models not yet employed in the marketing publications, including non-pre-trained models (FastText, CNN-LSTM, and BiLSTM) and pre-trained models recently developed in the computer science literature (BERT, XLNet, RoBERTa, DistilBERT, ALBERT), perform better than those previously applied in marketing (CNN and Feedforward NN (FNN)). F1-scores are about 61 on the top of Panel A for BERT and 49 at the bottom for KNN indicating about 20% improvement from bottom to top, and 50% improvement over the baseline "Frequent Class" method. Moreover, the pre-trained models preform particularly well. Although Panel B (sorted on F1) shows that CNN performs third, sixth, and eighth-best in the books, instant video, and automotive product categories, and FNN is seventh-best in the drug category, our first finding from Panel A applies largely to Panel B as well.

Although Panel C (sorted on MAE) shows that CNN performs sixth and eighth best in the books and automotive categories, and FNN is seventh best in the office category, our first finding from Panel A is found to be largely true in Panel C as well. In sum, Panels A, B, and C indicate that across the nine product categories, the neural network-based learning methods noted above (and particularly the newly developed pre-trained methods) are much more often found to appear in the top-5 performing (predictive) methods relative to those methods previously applied in marketing (CNN only appears once in the books category).

Second, as expected, Table 3 Panel A indicates most neural network-based learning methods perform better than other machine learning methods, some of which have been applied by marketing scholars (KNN, Naïve Bayes, SVM, Logistic Regression and Random Forest) and some of which (to our best knowledge) have not been applied by marketing scholars (Ordered Logistic Regression, Decision Tree, XGBoost and AdaBoost). While SVM and Logistic Regression have the fifth and ninth best F1 scores respectively, all other traditional machine learning methods perform less well. Confusion matrices (available from authors) for drugs, which have the most classes (10, relative to 3 and 5 for the other product categories) as an example explain how a better predictive model such as BERT outperforms KNN, Logit, Ordered Logit and XGBoost: because it has the highest correct predictions on the diagonal, relative to other approaches which appear to predict more extreme or most frequent ratings.

Panel B (sorted on F1) indicates that some traditional machine learning methods perform well, e.g., SVM is second-best in the patio and music categories and sixth-best in the hotel category; Ordered Logistic Regression is third-best in the musical instrument category; and XGBoost is fifth-best in the drug and office categories. However, our second finding of better predictions from neural network-based learning methods over other machine learning methods, observed in Panel A across product categories, is supported in Panel B in each product category.

In Panel C (sorted on MAE), Ordered Logistic Regression provides the best predictions in the patio and musical instrument categories, second-best in the office category, third-best in the automotive category, fourth-best in the instant video category, and fifth-best in the book category, potentially employing its suitability in fitting an ordinal sentiment rating. And SVM provides the third-best in the patio category, fifth best in the musical instrument category, and seventh-best in the hotel, airline, and automotive categories. However, once again, our second finding of better predictions from neural network-based learning methods over other machine learning methods, observed in Panels A across product categories and B in each product category based on F1, is also supported in Panel C in each product category (except patio and musical instrument products) based on MAE.

Third, in Panel A, the benchmark model that predicts all observations to be the most frequent class (sentiment rating), has an F1-score of 39.8 which is about 33% below the best F1-score of 61.0 for BERT. In other words, BERT provides a 50% improvement in F1-score over the most-frequent benchmark model. In Panel B, XLNet offers about 100%, 50%, and 10% improvements in the hotel, patio, and musical instrument categories, respectively. RoBERTa offers about 60% improvement in the airline category. BERT offers about 200%, 100%, 20%, and 50% improvements in the drug, book, automotive and instant video categories respectively. Consequently, the neural network-based learning models offer substantial improvements of F1 scores over the most-frequent benchmark model in each of the nine product categories. In addition, the Recall rates in Panel A range between 57% and 63% which are much higher than the random rates of 33%, 20% and 10% for 3-, 5- and 10-point scales respectively, and higher than the 55% Recall rate from the most-frequent benchmark model.

4.2. Diagnostic ability

In our setting, diagnostic ability is a data-based judgment that we as researchers will evaluate and make, based on the clarity of diagnostics afforded, i.e., the topics identified, and the words or lexical choices consumers employ to describe the topics identified. Our judgment is not unlike a judgment which can be made by any reader based on the data we provide.

We consider the diagnostic ability of three general approaches: (i) topic discovery models (such as LDA) which are explicitly designed for discovering words and topics, (ii) some non-neural network-based machine learning methods, such as logit and ordered logit, which are inherently capable of both diagnostics and predictions, and (iii) neural network models, which do not inherently have diagnostic ability but can be supplemented with newly developed tools (such as LIME) to provide simple diagnostics. We focus our discussion on the hotel dataset and airline dataset, each representing a general type of data source – review sites and social media. We discuss the three general approaches in turn.

Table 4
Diagnostics from Topic Discovery Applied to Data in Two Settings: Review Sites and Social Media.

Panel A. Hotels

#	Inferred Topic	Representative Words
1	Hotel Amenities	breakfast, clean, nice, good, great, parking, pool, staff, area, free, location
2	General Feedback	thank, feedback, hope, guest, experience, dear, time, appreciate, future, sincerely
3	Negative Experience	noise, bathroom, shower, door, water, bed, place, night, people, like, floor, small
4	Positive Experience	great, thank, enjoyed, staff, review, time, location, hear, soon, forward, happy
5	Booking	desk, told, called, check(ed), said, asked, night, booked, reservation, went
6	Room Layout	suite, bedroom, shuttle, kitchen, bed, king, living, large, floor, stayed, size, view
7	Location and Property	seattle, downtown, hyatt, home, chicago, great, house, staff, stayed, river, grand

Panel B. Airlines

#	Inferred Topic	Representative Words
1	Service Experience	thank(s), southwestair, united, seat(s), problems, booking, response, class
2	Payment	united, usairways, sure, bad, tell, customers, money, really, want, card, different, credit
3	Flight Cancelation/Delay	flight, cancelled, plane, united, gate, sitting, americanair, delayed, waiting
4	Budget Airline	jetblue, http, virginamerica, southwest(air), love, fleet, great, thanks
5	Reservation	hold, usairways, phone, help, americanair, need, change, trying, number, flight
6	Baggage Issues	service, customer, bag, united, airline, worst, baggage, lost, experience, terrible
7	Delay and Connection Issues	flight, late, usairways, americanair, time, delayed, miss, connection, thanks

Panel C. Drugs

#	Inferred Topic	Representative Words
1	Weight Control	weight, lost, pounds, lbs, appetite, hair, loss, started, eat, eating, gain, adhd, contrave
2	Skin	skin, face, using, burning, acne, product, itching, cream, use, used, yeast, red
3	Pain	pain, relief, works, sleep, years, muscle, patch, day, migraine, medication, medicine
4	Birth Control	birth, period(s), control, months, bleeding, mirena, insertion, spotting, got
5	Intimacy	period, pill, plan, took, days, later, sex, came, got, unprotected, condom, hours, day
6	Digestion	water, taste, day, drink, stomach, nausea, dose, diarrhea, prep, cough, infection
7	Anxiety	anxiety, feel/felt, life, depression, taking, effects, like, panic, years, zoloft, started

Panel D. Books

#	Inferred Topic	Representative Words
1	Family	life, mother, story, like, woman/women, novel, way, man, com, children, family
2	Positive Experience	read, books, reading, time, life, great, like, love, written, think, stories, recommend
3	Romance	love(d), story, read, series, life, great, characters, wait, new, family, romance
4	Feeling	love(d), like, read, really, series, books, story, know, good, characters, way, end
5	Overall Summary	story, read, really, like, character(s), good, reading, enjoyed, interesting
6	History	people, world, life, story, history, war, like, new, human, time, work, way, novel
7	Fantasy	character(s), world, story, novel, fantasy, read, time, magic, plot, series, like

Panel E. Automotive

#	Inferred Topic	Representative Words
1	Interior Cleaning	product, leather, car, cleaner, spray, clean, products, clay, cleaning, like, use(d)
2	Windshield and Light	wiper(s), light(s), blade(s), bulb(s), windshield, bright, rain, led
3	Fuel System	filter(s), oil, tank, price, hose, fuel, amazon, water, air, great, change, good, fits
4	Emission Control	oil, plug(s), fluid, engine, transmission, pump, drain, spark, cap, miles, horn
5	Tire	tire(s), gauge, battery, trailer, unit, use, works, easy, pressure, great, need, cord
6	Exterior Cleaning	wax, car, wash, battery, towel(s), charger, use, microfiber, paint, charge
7	Material	plastic, tape, tool, jeep, metal, fit, door, like, tight, place, little, cover, jack, hold

Panel F. Office

#	Inferred Topic	Representative Words
1	Writing Tools	pen(s), pencil(s), stapler, staple(s), ink, sharpener, writing, lead
2	Packing Tools	tape, binder(s), paper, cover, box, chair, scotch, plastic, dispenser, notebook
3	Filing Tools	folder(s), file(s), tab(s), desk, keyboard, monitor, filing, stand, calculator
4	Printing Devices	printer, print, printing, ink, epson, paper, cartridges, scanner, photo, quality, canon
5	Meeting Tools	board, phone(s), marker(s), erase, dry, magnets, eraser, magnetic
6	Notes	note(s), post, pad, sticky, mouse, wrist, surface, stick, adhesive, glue, use, like
7	Labeling	label(s), avery, template(s), address, cards, peel, use, print, sheet

Panel G. Patio

#	Inferred Topic	Representative Words
1	Feeding	feeder(s), water, deer, hummingbird(s), glass, nectar, clean
2	Plant	hose(s), plant(s), garden, soil, water, pot(s), grow, use, growing

3	Barbecue	grill, cover, weber, metal, nice, looks, sturdy, quality, plastic, fit, lid, box, pit, like
4	Rodent Control	trap(s), mouse/mice, feeder, bird(s), squirrel(s), bait, seed, set
5	Lawn Care	battery, mower, trimmer, gas, grass, power, use, lawn, handle, electric, cord, unit
6	Pest Control	product, ant(s), mole(s), charcoal, use(d), bugs, house, terro, killer, coals
7	Hand Tool	sprayer, use, spray, saw, product, branches, cut, chain, fiskars, scissors, cutting

Panel H. Music Instruments

#	Inferred Topic	Representative Words
1	Guitar Picks	pick(s), grip, dunlop, like, jazz, playing, feel, bass, medium, different
2	Audio Interface	cable(s), mic, recording, microphone, quality, audio, use, studio, usb, mixer
3	Guitar Pedal	amp, pedal(s), sound(s), tone, effects, distortion, like, power, volume
4	Guitar Straps	strap(s), guitar, leather, fit, screw(s), locks, planet, waves, hole(s)
5	Guitar Strings	string(s), guitar(s), capo, sound, acoustic, great, addario, good, play, tone
6	Gig Bag	stand, case, bag, guitar, stands, great, sturdy, good, price, works, fits, gig, music
7	Guitar Tuner	tuner(s), tune, tuning, snark, accurate, guitar, clip, instrument, ukulele

Panel I. Instant Videos

#	Inferred Topic	Representative Words
1	Comedy	comedy, funny, school, life, character, mother, lucy, plays, played, girl, role, family
2	Family	people, like, kids, watch, love, old, family, fun, watching, really, shows, children
3	Amazon Prime	like, good, characters, amazon, watch, really, series, shows, watching, pilot
4	Series	season(s), episode(s), series, new, doctor, wait, second, dvd, great
5	Actors	great, series, story, acting, good, excellent, actors, characters, love, recommend
6	Horror	movie(s), film(s), horror, story, like, director, minutes, way, end, really
7	History and Politics	world, war, series, new, agent, man, life, history, people, team, human, american

Panel J. Regression of Rating on Topic Proportions

Topics in Panels A-I	Hotels Rating	Airlines Rating	Drugs Rating	Books Rating	Automotive Rating	Office Rating	Patio Rating	Music Instruments Rating	Instant Videos Rating
Topic_1	4.621***	2.162***	7.628***	3.828***	4.397***	4.372***	4.017***	4.475***	3.840***
	(0.022)	(0.022)	(0.061)	(0.029)	(0.022)	(0.012)	(0.038)	(0.035)	(0.032)
Topic_2	3.767***	1.344***	6.596***	4.615***	4.257***	4.289***	4.224***	4.364***	4.254***
	(0.037)	(0.023)	(0.052)	(0.016)	(0.023)	(0.011)	(0.027)	(0.030)	(0.017)
Topic_3	2.246***	1.148***	7.943***	4.948***	4.740***	4.322***	4.400***	4.481***	4.039***
	(0.034)	(0.021)	(0.041)	(0.019)	(0.022)	(0.016)	(0.031)	(0.026)	(0.014)
Topic_4	5.114***	2.491***	6.186***	4.210***	4.610***	4.242***	3.827***	4.382***	4.821***
	(0.026)	(0.022)	(0.034)	(0.017)	(0.033)	(0.010)	(0.027)	(0.034)	(0.020)
Topic_5	2.103***	1.055***	7.912***	3.201***	4.566***	4.273***	4.457***	4.623***	4.835***
	(0.038)	(0.020)	(0.064)	(0.013)	(0.019)	(0.016)	(0.028)	(0.026)	(0.017)
Topic_6	4.676***	1.277***	6.342***	4.133***	4.525***	4.435***	3.985***	4.555***	2.802***
	(0.039)	(0.023)	(0.043)	(0.025)	(0.021)	(0.014)	(0.035)	(0.028)	(0.024)
Topic_7	5.091***	1.318***	6.850***	4.397***	4.241***	4.604***	4.327***	4.470***	4.278***
	(0.042)	(0.020)	(0.036)	(0.026)	(0.021)	(0.016)	(0.037)	(0.032)	(0.028)
Observations	10,000	14,640	62,348	39,160	20,467	53,237	13,258	10,254	37,125

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

4.2.1. Topic discovery

Although LDA cannot be directly applied for classification or prediction, it affords the best diagnostic ability, which allows us to analyze the content of reviews associated with different ratings. Table 4 shows the inferred seven topics for each of the nine product category datasets in Panels A through I and the associated words. In order to identify topics that discriminate better from worse ratings, Panel J reports regressions of ratings on the probability of a review being associated with each of the seven topics.[14] The decision to include seven topics for each of the nine product categories was made based on interpretability, i.e., based on whether the n + 1th topic offers qualitative interpretations which are different from the first n topics.

Why do we assess LDA as offering the best diagnostic ability? Because (i) LDA analyzes the interdependence between words and provides different groupings of interrelated words which an analyst can observe and label as separate inferred topics (like factor analysis which groups different variables to create separate factors). In addition, (ii) based on topics inferred from LDA, one can explore which inferred topics are associated with lower and higher levels of satisfaction, which leads to managerial implications.

For example, for the hotel dataset in Panel A, review topics 5 and 3, related to Booking (as represented by the words such as "desk", "told", "called", "check(ed)", "asked", "booked", and "reservation") and Negative Experience (as represented by the

[14] We choose not to include an intercept term because the seven topic probabilities add up to one, and multicollinearity means we would only obtain estimates for six of the seven probabilities if we were to include an intercept term. The results are more interpretable without an intercept term.

words such as "noise", "bathroom", "shower", "door", "water", and "small") are associated with lower levels of satisfaction or less positive coefficients of 2.10 and 2.25, respectively, in Panel J (recall the sentiment scale for hotels is 1–5). In contrast, topic 2, related to General Feedback (as represented by the words such as "thank", "feedback", "hope", "appreciate", "future", and "sincerely") is associated with a medium level of satisfaction, or a positive coefficient of 3.77 in Panel J. And, the topics associated with the highest levels of satisfaction include topic 4 which relates to Positive Experience (as represented by the words such as "great", "thank", "enjoyed", "staff", "forward", and "happy"), topic 7 which relates to Location and Property (as represented by the words such as "seattle", "downtown", "hyatt", "chicago", "river", and "grand"), topic 6 which relates to Room Layout (as represented by words such as "suite", "bedroom", "kitchen", "large", and "size"), and topic 1 which relates to Hotel Amenities (as represented by the words such as "breakfast", "clean", "parking", "pool", and "area"). Topics 4, 7, 6, and 1 have the highest positive coefficients of 5.11, 5.09, 4.68, and 4.62, respectively, in Panel J.

For the airline dataset compiled from Twitter, we find that the lowest satisfaction ratings in Panel J are associated with the coefficients 1.06, 1.15, 1.28, 1.32, and 1.34 (recall the sentiment scale for airlines is 1–3), which are associated with topics 5 (Reservation), 3 (Flight Cancelation/Delay), 6 (Baggage Issues), 7 (Delay and Connection Issues), and 2 (Payment), respectively (see Panel B). Reservation is associated with words such as "hold", "phone", "help", "trying", and "number". Flight Cancelation/Delay is represented by words such as "flight", "cancelled", "plane", "gate", "sitting", "delayed", and "waiting". Baggage Issues are represented by words such as "service", "customer", "bag", "worst", "baggage", and "lost". Delay and Connection Issues are associated with words such as "flight", "late", "time", "delayed", "miss", and "connection". Payment is associated with words such as "bad", "money", "card", "different", and "credit". In contrast, the highest coefficients in Panel J, 2.49 and 2.16, are respectively associated with topics 4 (Budget Airline) and 1 (Service Experience). Budget Airline is associated with words such as "jetblue", "virginamerica", and "southwest(air)", and Service Experience is associated with words such as "thank(s)", "seat(s)", "problems", "response", and "class".

Marketing practitioners can conduct such diagnostic analyses at either the product category level (like we have done) or the individual product or brand levels to diagnose the topics and related words inherent in the text-based reviews, which discriminate between better and worse consumer sentiment ratings.

4.2.2. Ordered logit

Our second main observation regarding diagnostic ability is that classic supervised non-neural network models such as ordered logit do not provide diagnostics as useful as those provided by LDA. Why? Because, while classification procedures such as ordered logit perform well and diagnose the determinant words or adjectives that are associated with positive versus negative sentiment, they do not provide the information on the association between the words. Consequently, it is not possible to understand the topics (issues) that the determinant words or adjectives are about (referring to).

For example, Table 5 presents the 30 most predictive words associated with positive and negative sentiment for hotel data (left) and airline data (right) from the ordered logit model. We choose to focus on the 30 most positive and negative words for a total of 60 words for each of the two datasets, because we identify about the same number of words for hotel and airline datasets using topic discovery, and we seek approximate comparability in the diagnostics achieved across methods. There are both similarities and differences between diagnostics provided by ordered logit (an analysis of dependence between numerical ratings and words) and LDA (an analysis of interdependence between words).

For example, for hotel data, positive words such as "clean" and negative words such as "told" are captured by both approaches. However, more single adjectives are identified by ordered logit to be positive words (e.g., "wonderful", "amazing", "excellent", "perfect", and "beautiful") or negative words (e.g., "terrible", "disgusting", "terrible", "worst", and "poor"), possibly because ordered logit analyzes dependence. However, because interdependencies between words are not identified as by LDA, it is not possible to identify the topics (e.g., hotel amenities, general feedback, negative experience, positive experience, room layout, and location and property) that the positive adjectives are associated with.

The two patterns observed for the hotel data are found for the airline data as well. Compared with LDA, ordered logit reveals more interesting single negative determinant words or adjectives, such as "nothing", "rude", "stuck", "disappointed", "ridiculous", and "poor", which are difficult to link to topics (e.g., Service Experience, Payment, Flight Cancellation/Delay, Budget Airline, Reservation, Baggage Issues, or Delay and Connection Issues), as is facilitated by LDA. Largely similar patterns are identified for each of the other seven product categories.

This potential shortcoming in the diagnostic ability of ordered logistic regression relative to topic discovery methods is because ordered logit is designed for prediction (not diagnostics) and operates at the word or lexical level, while topic discovery is designed for diagnostics (not prediction) and allows topic/issue-based words to be interdependent. With this noted, we do not find that positive (negative) words or adjectives have negative (positive) signs, which is a positive aspect.

4.2.3. Neural network models

It has traditionally been recognized that neural network models, while offering better predictions than non-neural network-based machine learning and topic discovery methods, are less capable of providing clear and concise diagnostics. The recent advances in explainable AI provide opportunities to achieve both predictive accuracy and diagnostic ability (Rai, 2020). Among the class of explainable AI techniques, LIME (Local[15] Interpretable Model-Agnostic Explanations), proposed

[15] "Local" refers to local fidelity that aims to imitate the behavior of the classifier nearby the instance.

Table 5
Thirty Most Determinant Words for Positive and Negative Sentiments obtained from Ordered Logit Model Applied to Data in Two Settings: Review Sites and Social Media.

Review Site (Hotel Dataset)				Social Media (Twitter Airline Dataset)			
Positive Words	Positive Coefficients	Negative Words	Negative Coefficients	Positive Words	Positive Coefficients	Negative Words	Negative Coefficients
great	5.5601	dirty	−4.9168	thank	8.3504	hour	−6.0004
wonderful	4.2996	sorry	−4.8554	great	5.6670	delay	−4.6621
amazing	4.0319	rude	−3.9726	awesome	5.2121	worst	−4.6537
excellent	3.9535	bad	−3.8690	amazing	5.0877	hold	−3.8154
perfect	3.6340	apologize	−3.8662	love	4.5652	luggage	−3.3801
love	3.5522	horrible	−3.6420	best	4.0106	cancelled	−3.3120
beautiful	3.5238	disgusting	−3.2705	appreciate	3.4437	bag	−3.2816
awesome	3.1697	told	−3.2434	kudos	3.4275	nothing	−2.9402
clean	2.9201	terrible	−3.2213	virginamerica	3.0030	late	−2.9267
always	2.7551	loud	−3.2150	worries	2.8639	rude	−2.9071
friendly	2.6070	management	−3.0511	jetblue	2.6586	lost	−2.9015
emma	2.5327	smelled	−2.9692	good	2.5554	stuck	−2.8685
spacious	2.4947	worst	−2.9594	wonderful	2.4441	disappointed	−2.7612
husband	2.4747	booked	−2.9343	southwestair	2.4152	system	−2.6855
help	2.4421	filthy	−2.9341	excellent	2.3654	customers	−2.6733
everything	2.3737	stain	−2.8545	enjoy	2.3244	call	−2.6400
outstanding	2.3671	poor	−2.7937	favorite	2.3065	website	−2.6018
comfortable	2.3291	work	−2.7648	rock	2.2445	paid	−2.5361
every	2.2948	cockroaches	−2.7125	glad	2.1489	never	−2.5069
highly	2.2813	charged	−2.6875	welcome	2.1319	terrible	−2.3949
accommodating	2.2590	smell	−2.4621	comfortable	2.0466	fail	−2.3837
suite	2.1932	sheets	−2.4343	cool	2.0458	still	−2.2821
enjoy	2.1057	would	−2.4216	warm	2.0003	days	−2.2459
fantastic	2.0733	unfortunately	−2.3912	definitely	1.9786	waiting	−2.1891
professional	2.0374	walls	−2.3295	haha	1.9709	ridiculous	−2.1223
smile	2.0102	nothing	−2.2852	worked	1.9671	hung	−2.1205
return	2.0057	ok	−2.2544	appreciated	1.8807	pay	−2.1198
relaxing	1.9537	falling	−2.2004	outstanding	1.8726	trying	−2.0883
fabulous	1.9506	mold	−2.1819	nice	1.8717	stranded	−2.0692
best	1.9348	money	−2.1817	impressed	1.8599	poor	−2.0416

by Ribeiro et al. (2016), can be applied to provide model-agnostic local explanation to neural network models. This procedure is model-agnostic because it works on any classifier. Its explanation could hypothetically be interpretable by the researcher.

We apply LIME on BERT, one of the methods with the best predictive ability, to obtain diagnostics. Because LIME generates explanations for specific predictions, we use all the instances in each test set of the hotel and airline datasets and aggregate the explanations to assess whether LIME can distinguish between positive and negative words. Details on the approach are provided in Online Appendix B.

We use Pointwise mutual information (PMI), a widely used metric in NLP, to compute associations between words that consumers use to describe their product experience, thereby identifying importance of each attribute (Church & Hanks, 1990). Table 6 shows PMI of explained words for the hotel (left) and airline (right) datasets. For the hotel dataset, words such as "friendly", "excellent", "enjoy", "beautiful", and "fantastic" have high PMI for positive sentiments, while words such as "rude", "dirty", "worse", "awful", and "terrible" have high PMI for negative sentiments. There is much consistency in the positive and negative words across Table 5 on ordered logistic diagnostics and Table 6 on BERT. Like the ordered logit diagnostics in Table 5, our third main observation on diagnostic ability is that it is not possible for the positive and negative words identified by BERT and LIME in Table 6 to be linked to the topics, but positive (negative) adjectives, in large part, do not have negative (positive) PMIs.

Our third main observation on diagnostic ability based on the hotel dataset is observed for the airline dataset as well, in that words such as "great", "best", "new", "happy", and "amazing" have high PMI for the positive sentiments, while words such as "cancelled", "never", "lost", "stuck", and "late" have high PMI for the negative sentiments. There is also much consistency between the words across Tables 5 and 6, as we observed for ordered logit diagnostics. However again, it is difficult to infer whether each positive and negative word is associated with one or another topic, an aspect which is possible when topic discovery methods such as LDA are employed. In addition, conceptual gradations between positive adjectives such as "great", "amazing", "best", "wonderful", and "enjoy", as well as between negative words such as "delay", "cancelled", "late", "lost", and "stuck" are difficult but potentially mitigated by PMI values. Largely similar results are obtained for each of the other product categories.

Fig. 1 Panels A and B provide examples of LIME explanations for hotels and airlines respectively. Essentially, LIME identifies words associated or not associated with each outcome (the numerical rating in our case). For example, LIME can identify the words associated with a rating of 3 and not 3, 5 and not 5, or 10 and not 10, for airlines, hotels (and six other product categories), and drugs respectively. LIME can report reviews which provide information about how the words are employed

Table 6
Thirty Most Determinant Words for Positive and Negative Sentiments obtained from BERT and LIME Applied to Data in Two Settings: Review Sites and Social Media.

Review Site (Hotel Dataset)				Social Media (Twitter Airline Dataset)			
Positive Word	Positive PMI	Negative Word	Negative PMI	Positive Word	Positive PMI	Negative Word	Negative PMI
great	0.2561	told	2.5376	thank	1.8920	hour	0.4424
friendly	0.2554	booked	2.5103	jetblue	1.8875	cancelled	0.4415
love	0.2549	check	2.4982	southwestair	1.8863	delay	0.4389
excellent	0.2545	inconvenient	2.4932	good	1.8739	hold	0.4352
everything	0.2542	said	2.4932	great	1.8735	wait	0.4346
always	0.2536	rude	2.4875	virginamerica	1.8735	still	0.4336
enjoy	0.2523	door	2.4811	love	1.8599	bag	0.4271
perfect	0.2521	smell	2.4737	best	1.8408	never	0.4254
beautiful	0.2462	literally	2.4652	nice	1.8408	get	0.4245
spacious	0.2445	poor	2.4494	appreciate	1.8385	usairways	0.4238
wonderful	0.2428	dirty	2.4431	new	1.8385	say	0.4218
highly	0.2419	someone	2.4431	http	1.8270	told	0.4199
stay	0.2415	paid	2.4286	well	1.8270	phone	0.4199
awesome	0.2404	tired	2.4286	awesome	1.8194	minutes	0.4192
fantastic	0.2388	worse	2.4286	happy	1.7979	lost	0.4178
super	0.2374	gross	2.4107	helpful	1.7905	call	0.4162
helpful	0.2348	awful	2.4107	amazing	1.7819	even	0.4144
amazing	0.2325	moldy	2.4107	yes	1.7819	stuck	0.4135
everyone	0.2274	disappointing	2.4107	tonight	1.7719	problems	0.4114
modern	0.2274	looked	2.4001	tomorrow	1.7719	trying	0.4103
best	0.2264	terrible	2.4001	wow	1.7599	luggage	0.4091
hyatt	0.2254	sorry	2.3942	platinum	1.7599	time	0.4072
delighted	0.2254	saying	2.3880	forward	1.7599	customers	0.4065
accommodating	0.2254	cockroaches	2.3880	dfw	1.7454	worst	0.4065
comfortable	0.2254	garbage	2.3880	wish	1.7454	late	0.4044
happy	0.2221	one	2.3880	please	1.7454	bad	0.4036
nice	0.2218	sheets	2.3880	glad	1.7275	someone	0.3985
exceptional	0.2195	got	2.3880	first	1.7275	nothing	0.3965
close	0.2181	stains	2.3880	oh	1.7275	sitting	0.3965
seattle	0.2181	towels	2.3880	big	1.7275	last	0.3943

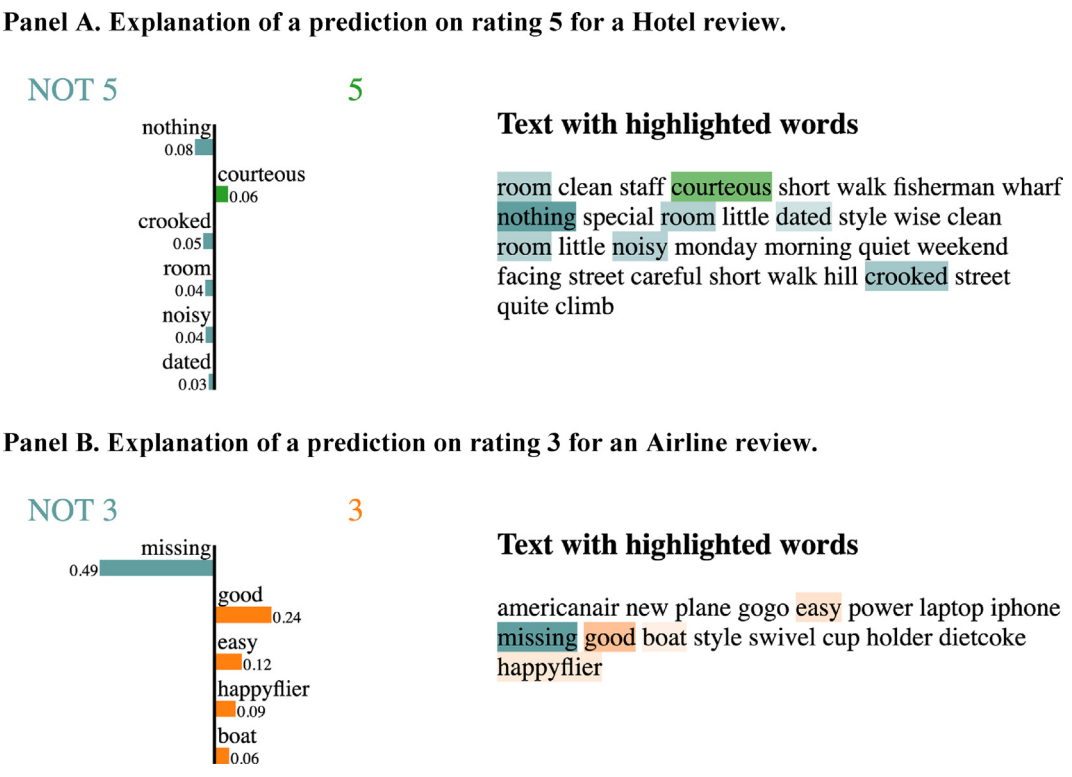

Fig. 1. Examples of LIME Multiclass Explanations of BERT in Two Settings: Review Sites and Social Media.

together. For example, in Panel A for hotels, the word associated with a single prediction of the rating 5 is "courteous", and words associated with that single prediction which is not 5 are "nothing", "crooked", "room", "noisy", and "dated". Similarly, in Panel B for airlines, the words associated with a single prediction of the rating 3 are "good", "easy", "happyflier", and "boat", while the word associated with that single prediction which is not 3 is "missing".

However, to infer the topics as in LDA, many such reviews would need to be read and further processed by analysts and managers. Therefore, LIME is a potential approach to mitigate the disadvantage of neural network models in lack of diagnostic ability, but the resulting diagnostics are restricted by a few limitations. First, it approximates the model, so may not be entirely accurate. Second, it works at each prediction level, and therefore it is difficult to diagnose the whole classifier. Third, neural network models are focused on predictions and ignore correlations that are not helpful for predictions, thus their diagnostics will always be incomplete. In this regard, the diagnostic ability of neural network models is inferior to that of topic discovery models. While LIME does not solve the diagnostic problem of neural network models, it provides a promising direction towards interpretation.

5. Conclusion and discussion

Undoubtedly, the analysis of unstructured data, especially user-generated texts, has vast yet undiscovered potential (Berger et al., 2020; Humphreys & Wang, 2018). Through our comprehensive review of techniques that can be utilized to analyze consumer review data, we seek to inspire, inform, and guide marketing scholars, analysts and managers, in selecting methods for future applications.

We focus on predictive and diagnostic abilities to assess the performance of different tools in addressing the fundamental relationship between the numerical rating and the review text. In terms of predictive ability, we find that the recently proposed pre-trained neural network methods in the computer science literature (e.g., BERT, XLNet, RoBERTa, DistilBERT, and ALBERT), and several non-pre-trained neural network methods (e.g., FastText, CNN-LSTM, and BiLSTM) provide better predictions than those previously employed in the marketing publications. Although the improved predictive ability requires additional efforts on hyper-parameter tuning,[16] this limitation will reduce with enhancement in computing power.

Why is this finding on predictive ability important? Which marketing stakeholders can utilize the finding? And how can they utilize the finding? Marketing science scholars have a long-standing tradition of evaluating the predictive abilities of competing customer preference, choice, or market response models, and choosing the model with the highest predictive ability for their discussions of diagnostics. These diagnostics are typically based on parameter estimates and focus on the effects of product attributes, marketing-mix efforts, and customer characteristics on customer decisions and market outcomes (Abramson et al. 2000; Andrews, Ainslie, et al., 2002; Andrews, Ansari, et al., 2002; Andrews et al., 2008, 2011). Our findings on predictive ability suggest that such scholars will want to employ the pre-trained neural network models over their non-pre-trained variants because of their better predictive abilities. In addition, practitioners who seek to accurately identify customers who have more positive sentiment and preferred behaviors over others, are likely to prefer the pre-trained neural network models over their non-pre-trained counterparts for similar reasons.

In terms of diagnostic ability, our main finding is that the neural network models which are best on prediction do not provide the best diagnostics under the current state of technology. Although these methods generate several determinant words and adjectives related to the product usage experience, it is difficult to associate the adjectives to topics underlying the experience. In contrast, the topic discovery method (LDA), which is incapable of prediction, permits an association of the adjectives with topics, so that it is possible for analysts, managers, and scholars to understand how consumer sentiment can be improved in product usage settings.

Why is this finding on diagnostic ability important? Which marketing stakeholders can utilize the finding? And how can they utilize the finding? Marketing practitioners are often looking for ways in which the determinant words and topics from sentiment models can quickly provide useful low-cost managerial inputs aimed at improving the design and marketing of products and services, and the corresponding consumer experience. For this purpose, because of reasons mentioned in the previous paragraph, we assess LDA as preferable to the classification procedures.

For example, in Table 4, on hotels (Panel A) practitioners can quickly learn that sentiment is driven by seven topics and examine each of the seven topics. For example, one of the topics is Hotel Amenities and that this topic is associated with generally positive consumer sentiment (Column 1, Panel J). Thus, the top words under this topic (breakfast, parking, pool, and staff) suggest four foci for practitioners' efforts to provide a clean and positive experience. Similar insights are available for each of the six other topics. Similarly, for airlines (Panel B), practitioners can quickly learn about seven drivers of sentiment, one of which is Flight Cancellations/Delay which is associated with generally negative sentiment (Column 2, Panel J). The representative words associated with Flight Cancellation/Delay (e.g., sitting, waiting, gates) suggest foci for practitioners' efforts to provide refreshments, rationale for the delay, and forward-looking information and updates aimed at mitigating the negative experience. Similar insights are available for each of the six other topics.

The implication of our main findings on predictive and diagnostic abilities is that there is no "winner takes all", no one method is best for all tasks, much like all other marketing methodological approaches prior to sentiment analysis (focus

[16] To reduce the training time for neural network models, we suggest a computer with a Graphics Processing Unit (GPU) supporting packages such as TensorFlow and PyTorch.

groups, surveys, models, and experiments). Consequently, method choices in future applications based on text review data can and will vary based on the goal in marketing applications, whether it is predictive, e.g., for targeting purposes, or diagnostic, e.g., for improving customer sentiment, consumer perceptual mapping and creating product differentiation maps.

If the analyst seeks to strike a balance, s/he can engage in a two-step modeling process: (i) summarize the topics of the text data with LDA, and (ii) leverage the rating data to estimate more complex models of consumers' consideration, purchase decisions, and the overall experience. Such a hybrid modeling process would result in some loss of predictive ability but can potentially enable diagnostics. Finally, dictionary-based methods will continue to be employed by researchers who are interested in identifying and scaling emotional, and more generally certain psychological, sociological, or anthropological theory-based constructs which are pre-defined in dictionaries.[17]

Every work has limitations which afford avenues for future research. First, this paper focuses on the fundamental relationship between the overall numerical sentiment rating and the text-based explanation of that sentiment or rating. There are other independent variables which will influence numerical ratings, and there are dependent variables other than numerical ratings which an analyst may be interested in. Second, we focus on review text from two settings, review sites (e.g., Amazon, Goodreads) and social media (e.g., Twitter). However, text data, more generally, are available in many forms, such as text messages, emails, posts and blogs on online forums, and feeds on social networking sites. These data can be employed for a variety of purposes such as to study opinion leadership, development of consumer communities, social media firestorms, and word-of-mouth communications (Humphreys & Wang, 2018). Third, it is a current challenge to make neural network models capable of providing diagnostics from the same model, labeled as Explainable AI or interpretable machine learning (Rai, 2020), which represents tremendous opportunities for future research. We hope our work will help marketing researchers, practitioners and scholars select appropriate tools for analyzing text review data based on their goals (predictive, diagnostic) and that future research builds on our efforts in directions identified.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2021.10.011.

References

Abramson, C., Andrews, R. L., Currim, I. S., & Jones, M. (2000). Parameter bias from unobserved effects in the multinomial logit model of consumer choice. *Journal of Marketing Research, 37*(4), 410–426.

Anderson, E. T., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research, 51*(3), 249–269.

Andrews, R. L., Ainslie, A., & Currim, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research, 39*(4), 479–487.

Andrews, R. L., Ainslie, A., & Currim, I. S. (2008). On the recoverability of choice behaviors with random coefficients choice models in the context of limited data and unobserved effects. *Management Science, 54*(1), 83–99.

Andrews, R. L., Ansari, A., & Currim, I. S. (2002). Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *Journal of Marketing Research, 39*(1), 87–98.

Andrews, R. L., Currim, I. S., & Leeflang, P. S. H. (2011). A comparison of sales response predictions from demand models applied to store-level versus panel data. *Journal of Business & Economic Statistics, 29*(2), 319–326.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science, 57*(8), 1485–1509.

Bai, X., Marsden, J. R., Ross Jr, W. T., & Wang, G. (2020). A note on the impact of daily deals on local retailers' online reputation: Mediation effects of the consumer experience. *Information Systems Research, 31*(4), 1132–1143.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science, 323*(5919), 1297–1298.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing, 84*(1), 1–25.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science, 35*(6), 953–975.

Chakraborty, I., Kim, M., & Sudhir, K. (2021). *Attribute sentiment scoring with online text reviews: Accounting for language structure and attribute self-selection. Cowles Foundation Discussion Paper, No. 2176R2*.

Chen, Z. (2017). Social acceptance and word of mouth: How the motive to belong leads to divergent WOM with strangers and friends. *Journal of Consumer Research, 44*(3), 613–632.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Fishbein, M., & Ajzen, I. (1977). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Gensch, D. H., & Recker, W. W. (1979). The multinomial, multiattribute logit choice model. *Journal of Marketing Research, 16*(1), 124–132.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science, 31*(3), 493–520.

Ghose, A., Ipeirotis, P. G., & Li, B. (2019). Modeling consumer footprints on search engines: An interplay with social media. *Management Science, 65*(3), 1363–1385.

Goes, P. B., Lin, M., & Au Yeung, C. (2014). "Popularity effect" in user-generated content: Evidence from online product reviews. *Information Systems Research, 25*(2), 222–238.

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). In *Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning*, Proceedings of the 2018 International Conference on Digital Health. 2018. .

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research, 5*(2), 103–123.

Harbert, T. (2021). *Tapping the power of unstructured data*. https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data..

[17] For some research questions, the word list/expression list is finite and therefore sophisticated methods might not be necessary (for example mining for the number of personal pronouns to measure the scale of involvement or classify texts).

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing, 36*(1), 20–38.
Heitmann, M., Siebert, C., Hartmann, J., & Schamp, C. (2020). *More than a feeling: Benchmarks for sentiment analysis accuracy*. Available at SSRN 3489963.
He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web*, 507–517.
Humphreys, A., & Wang, R. J. H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research, 44*(6), 1274–1306.
Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 638–646).
Kirmani, A., Hamilton, R. W., Thompson, D. V., & Lantzy, S. (2017). Doing well versus doing good: The differential effect of underdog positioning on moral and competent service providers. *Journal of Marketing, 81*(1), 103–117.
Kübler, R. V., Colicev, A., & Pauwels, K. H. (2020). Social Media's Impact on the Consumer Mindset: When to Use Which Sentiment Extraction Tool?. *Journal of Interactive Marketing, 50*, 136–155.
Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research, 48*(5), 881–894.
Liu, X., Lee, D., & Srinivasan, K. (2019). Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research, 56*(6), 918–943.
Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing, 77*(1), 87–103.
McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 43–52.
Packard, G., & Berger, J. (2017). How language shapes word of mouth's impact. *Journal of Marketing Research, 54*(4), 572–588.
Puranam, D., Narayan, V., & Kadiyali, V. (2017). The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors. *Marketing Science, 36*(5), 726–746.
Rafieian, O., & Yoganarasimhan, H. (2021). Targeting and privacy in mobile advertising. *Marketing Science, 40*(2), 193–218.
Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141.
Ransbotham, S., Lurie, N. H., & Liu, H. (2019). Creation and consumption of mobile word of mouth: How are mobile reviews different?. *Marketing Science, 38*(5), 773–792.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).
Rubera, G. (2015). Design innovativeness and product sales' evolution. *Marketing Science, 34*(1), 98–115.
Sridhar, S., & Srinivasan, R. (2012). Social influence effects in online product ratings. *Journal of Marketing, 76*(5), 70–88.
Steckel, J. H., & Vanhonacker, W. R. (1993). Cross-validating regression models in marketing research. *Marketing Science, 12*(4), 415–427.
Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*(1), 1–20.
Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science, 31*(2), 198–215.
Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research, 51*(4), 463–479.
Tirunillai, S., & Tellis, G. J. (2017). Does offline TV advertising affect online chatter? Quasi-experimental analysis using synthetic control. *Marketing Science, 36*(6), 862–878.
Van Laer, T., Edson Escalas, J., Ludwig, S., & Van Den Hende, E. A. (2019). What happens in Vegas stays on TripAdvisor? A theory and technique to understand narrativity in consumer reviews. *Journal of Consumer Research, 46*(2), 267–285.
Vermeer, S. A. M., Araujo, T., Bernritter, S. F., & van Noort, G. (2019). Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media. *International Journal of Research in Marketing, 36*(3), 492–508.
Villarroel Ordenes, F., Ludwig, S., De Ruyter, K., Grewal, D., & Wetzels, M. (2017). Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. *Journal of Consumer Research, 43*(6), 875–894.
Wan, M., & McAuley, J. (2018). Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 86–94).
Wan, M., Misra, R., Nakashole, N., & McAuley, J. (2019). Fine-grained spoiler detection from large-scale review corpora. *ArXiv Preprint*. ArXiv:1905.13416.
Wang, Y., & Chaudhry, A. (2018). When and how managers' responses to online reviews affect subsequent reviews. *Journal of Marketing Research, 55*(2), 163–177.
Wu, C. (2015). Matching value and market design in online advertising networks: An empirical analysis. *Marketing Science, 34*(6), 906–921.
Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science, 66*(3), 1045–1070.
Zhang, Y., & Godes, D. (2018). Learning from online social ties. *Marketing Science, 37*(3), 425–444.