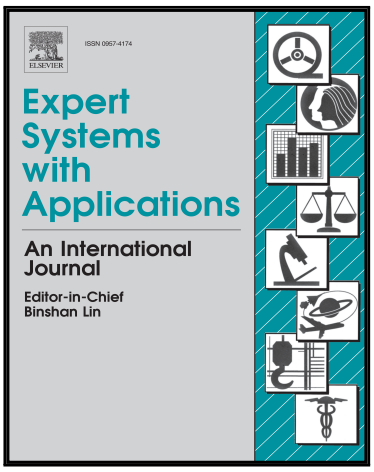

Predicting the Helpfulness of Online Reviews Using a Scripts-Enriched Text Regression Model

Thomas L. Ngo-Ye , Atish P. Sinha , Arun Sen

PII: S0957-4174(16)30664-9
DOI: 10.1016/j.eswa.2016.11.029
Reference: ESWA 11000

To appear in: *Expert Systems With Applications*

Received date: 3 July 2016
Revised date: 19 November 2016
Accepted date: 20 November 2016

Please cite this article as: Thomas L. Ngo-Ye , Atish P. Sinha , Arun Sen , Predicting the Helpfulness of Online Reviews Using a Scripts-Enriched Text Regression Model, *Expert Systems With Applications* (2016), doi: 10.1016/j.eswa.2016.11.029

Predicting the Helpfulness of Online Reviews Using a Scripts-Enriched Text Regression Model

Thomas L. Ngo-Ye[a,*]
[a]Department of Computer Information Systems
College of Business Administration
Alabama State University
Montgomery, AL 36101
U.S.A.
tngoye@alasu.edu

Atish P. Sinha[b]
[b]Lubar School of Business
University of Wisconsin-Milwaukee
Milwaukee, WI 53201-0742
U.S.A.
sinha@uwm.edu

Arun Sen[c]
[c]Mays Business School
Texas A&M University
College Station, Texas 77843
U.S.A.
arunsen.tx@gmail.com

[*]Corresponding author. Tel.: +1 334 229 5729

November 19, 2016

ABSTRACT

In this paper, we examine the utility of script analysis for predicting the helpfulness of online customer reviews. We employ the lens of cognitive scripts and posit that people share a cognitive script for what constitutes a helpful review in a given domain. Conceptually, a script includes the salient elements that readers look for before determining whether a review is helpful. To operationalize the construct of cognitive script, we seek the help of human annotators and ask them to highlight phrases that they believe are important for determining review helpfulness. The words in the annotated phrases are collected and become part of the script lexicon for a given domain. The lexicon entries represent the shared conception of essential elements, which are key to the evaluation of review helpfulness. We employ the words in the script lexicon as features in a text regression model to predict review helpfulness. Furthermore, we develop and empirically validate a new approach for combining script analysis and dimension reduction. The purpose of the study is to propose a new method to predict review helpfulness and to evaluate the effectiveness and efficiency of the scripts-enriched model. To demonstrate the efficacy of the scripts-enriched model, we compare it with benchmark models – a Baseline model and a bag-of-words (BOW) model. The results show that the scripts-enriched text regression model not only produces the highest accuracy, but also the lowest training, testing, and feature selection times.

Keywords: Business intelligence, Online customer reviews, Review helpfulness, Script theory, Human annotation, Text regression

1. Introduction

The Internet has become a popular venue for exchanging ideas, opinions, and views on almost every imaginable subject (Chen & Zimbra, 2010). Among them, online customer reviews on products and services are particularly valuable for consumers seeking independent opinions. Online reviews are also important to manufacturers and service providers, as the reviews may reveal customers' genuine concerns and provide useful market intelligence (Huang & Korfiatis, 2015). They can also be employed to investigate the effectiveness of product differentiation strategies (Clemons, Gao, & Hitt, 2006).

What makes online customer reviews helpful to a potential buyer in the process of making a purchase decision is an important research question (Mudambi & Schuff, 2010). For some popular products, there are hundreds or even thousands of online customer reviews. Going through all of them is extremely time-consuming for both consumers and businesses. Moreover, not all online reviews are equally helpful. Why some reviews have no or few helpful votes, while others obtain many useful votes, is an interesting research question (Cao, Duan, & Gan, 2011).

Ensuring and encouraging the quality of user-generated content in online communities is a real challenge (Chen, Xu, & Whinston, 2011). To help consumers manage the information overload problem in the online review domain, many review websites adopt the human feedback mechanism to sort and rank reviews, the so-called "social navigation" method. Major web sites such as Amazon.com and Yelp.com provide a "Most Helpful First" option in sorting and presenting customer reviews. These websites usually ask the question: "Was this review helpful? (Yes or No)". Based on the number of readers' votes, the customer reviews are ranked on the "helpfulness" dimension.

In this research, we investigate the problem of predicting online review helpfulness by employing the lens of *concept scripts* (Schank, 1982; Schank, 1999; Schank & Abelson, 1977). The core principle of script theory is that cognitive scripts represent shared human knowledge of a specific domain. We design a human annotation study to elicit a *script lexicon* of helpful online reviews and incorporate it as a new feature in script models to predict review helpfulness. By exploiting the high-level memory structures of

scripts, we expect to be able to better estimate review helpfulness than in situations when human knowledge is absent.

We collect real-world reviews and conduct an empirical study to compare script-based models with bag-of-words (BOW) models, as well as with a Baseline model that includes metadata features such as review depth (length), review readability, review sentiment, review valence (star rating), review extremity, and review age. We find that the Script models outperform both the Baseline and the BOW models. The empirical findings demonstrate the effectiveness of scripts for predicting review helpfulness. This research demonstrates a new usage of cognitive script – employing annotated script terms, which represent shared domain knowledge – as new features in text regression models for estimation of review helpfulness. This study is also the first to extract users' cognitive scripts in the context of online customer reviews. Moreover, the proposed novel script theory-based approach provides an attractive alternative to the computationally expensive text mining approach.

2. Theoretical background

2.1. Previous research on review helpfulness

In recent years, review helpfulness has been the focus of several studies. Some researches examine both review textual content and reviewer characteristics. For example, Ngo-Ye and Sinha (2014) study the influence of reviewer engagement characteristics, such as reputation, commitment, and current activity, on online review helpfulness. They find that reviewer engagement characteristics further enhance the review usefulness prediction beyond review text. Zheng, Zhua, and Lin (2013) also demonstrate that reviewers' social characteristics are important in improving review quality classification results. Using the dual process theory, Baek, Ahn, and Choi (2013) classify factors influencing the helpfulness of reviews into central cues for systematic information processing and peripheral cues for heuristic information processing. They discover that both peripheral cues (reviewer's credibility and review rating) and central cues (content of reviews) affect review helpfulness. Similarly, Li, Huang, Tan, and Wei (2013) show that both source-based and content-based review features affect review helpfulness.

Moreover, their empirical results indicate that concrete reviews are regarded as more helpful than abstract reviews by consumers. Zhu, Yin, and He (2014) study the direct impact of reviewer credibility on review helpfulness and find that the number of friends of a reviewer (online attractiveness) and the number of "Elite" badges a reviewer receives (expertise) both contribute helpfulness votes for reviews. However, reviews authored by an opinion leader (a reviewer with more online friends and more Elite badges) do not necessarily obtain more usefulness votes (Zhu, Yin, & He, 2014).

Review sentiment, rating extremity, length, and longevity have also been investigated. Using a sentiment mining approach to big data analytics, Salehan and Kim (2016) discover that sentimental reviews with neutral polarity in text are considered as more helpful. Review length and longevity also positively affect review helpfulness (Salehan & Kim, 2016). Baek, Lee, Oh, and Ahn (2015) investigate the relationship between review rating extremity and review helpfulness. They find that when a review's rating is close to the average rating of the review object, the review is regarded as helpful, indicating the role of normative social influence. Similarly, Yin, Mitra, and Zhang (2016) provide empirical support for confirmation bias, where readers tend to regard reviews that confirm their initial beliefs as more useful. They show that when the average product rating is high, positive reviews are considered more useful.

Some recent studies have pointed out that the review helpfulness ratings based on readers' votes are biased. For example, Wan and Nakayama (2014) examine the reliability of online review helpfulness and demonstrate that helpfulness ratings for most helpful reviews are inflated due to online consumers' self-selection behavior. Moreover, Kuan, Hui, Prasarnphanich, and Lai (2015) evaluate the impact of a set of review characteristics (review length, readability, valence, extremity, and reviewer credibility) on review helpfulness and review voting. Due to sample selection (review voting) bias, review helpfulness can be over-estimated (when review characteristics have opposite effects on review voting and review helpfulness) or under-estimated (when the effects of review characteristics on review voting and review helpfulness are in the same direction) (Kuan, Hui, Prasarnphanich, & Lai, 2015).

2.2. Cognitive script theory

Schank and Abelson (1977) developed the Cognitive Script Theory. A concept script encapsulates normal expectations about an activity. A concept script is like an outline of a play and is an abstraction of a routine activity (Schank & Abelson, 1977; Rumelhart & Norman, 1988). The format of a common activity, such as an outing at a restaurant, can be organized and abstracted in the form of a concept script (Schank & Abelson, 1977; Rumelhart & Norman, 1988). The scripts for many common activities are, to a large extent, well shared and agreed upon among people in a culture (Bower, Black, & Turner, 1979). As a fundamental mental organization tool, scripts are pervasive in people's cognitive activities. Scripts capture the specific human knowledge that characterizes an event or an activity. Cognitive Script Theory (Schank, 1982; Schank, 1999) maintains that when processing new inputs, people search through their existing memory to locate the most approximate object/event and form new knowledge based on their past experience. In other words, people understand and learn new things with reference to past experience and knowledge stored in memory. It is through this reminding process that the existing mental script gets reinforced. With the mechanism of reminding and learning (Schank, 1982), scripts help people form and fulfill expectations and reinforce past beliefs. Schank and Abelson (1977) suggest that higher level knowledge structures are helpful for understanding text. In their original conceptualization, they argue that higher level knowledge structures, such as scripts, supply background information and help people making assumptions and inferences while processing the text.

Leigh and Rethans (1984) conduct an empirical study of script-theoretic analysis of industrial purchasing behavior. Following a free elicitation procedure (Bower, Black, & Turner, 1979), Leigh and Rethans (1984) solicit 36 industrial buyers to complete a self-administered questionnaire eliciting concept scripts related to a computer terminal purchase. For script elicitation, they employ basic counts and basic descriptive statistics of specific activities or events. The presence of detailed activities in scripts is consistent with the view that cognitive scripts comprise fairly concrete episodic experiences as the basic building blocks (Abelson, 1976). Script theory has also been used in IS research. Multiple experts were interviewed to freely elicit scripts. A script lexicon and construct knowledge decision tree model were

developed to show the logical sequence of selling activities (Ainscough, DeCarlo, & Leigh, 1996). Script theory-based analysis is described as one of methods in the text analysis framework (Lacity & Janson, 1994). In the context of text analysis, the understanding of the meaning is derived from making sense of the non-random patterns in the text collection, such as the frequency or percentage of interested concepts and their relationships. After quantifying the count measurement via scheme coding or scripts categorization, traditional statistical hypothesis tests can be performed to validate the proposed theory (Lacity & Janson, 1994). For example, Barley (1986; 1990) found the non-random pattern that the frequency of recurring scripts was significantly correlated with the extent of new technology induced decentralization.

3. Conceptual model development

In this section, we first describe how we can apply script theory to online customer reviews. We propose a new method for generating script lexicons for online reviews. We then describe the development of conceptual models.

3.1. Script analysis of online customer reviews

Scripts are present in various offline consumer decision making scenarios (Erasmus, Boshoff, & Rousseau, 2002). In the online consumer decision scenario, the text in a customer review is a manifestation of a reviewer's personal experience with a given object (e.g., a digital camera) or event (e.g., dining at a restaurant). Thus, an online review is the reflection of a person's particular mental reconstruction of his/her direct experience with the reviewed object. On the other hand, a concept script is a mental structure of knowledge organization, a cluster of shared knowledge about the focus point (Weiten, 2008). Online review text and a concept script are associated because of their shared cognitive aspect. While the text of a review represents a specific cognitive processing by a single reviewer, a concept script represents collective cognitive processing. According to the principle of reminding from cognitive script theory, the event of reading the text of a review reminds him/her of an existing mental representation (script) of certain events and objects. A script also guides people to fulfill expectations

while processing text information. Text that resonates with a script tends to draw more cognitive attention and strengthen beliefs.

A reader has a mental script, which captures his/her expectations in terms of what he/she wants to see in a review. Some reviews contain elements that are relevant to the script. Such reviews would be considered to be useful. Other reviews, which fail to address the relevant elements, do not help the readers to update their beliefs and, therefore, are not regarded as helpful. Reviews covering salient elements of a mental script would be more effective in satisfying readers' information needs. Because scripts facilitate readers' understanding, assimilation, and learning (Schank, 1982), a review text that contains many elements of the underlying script will likely be regarded as more helpful by readers.

In traditional script theory-based applications, human subjects are often asked to write down the knowledge in their minds in an open-ended format. Researchers count the most frequent items and assemble them into a script lexicon, which contains common items above a certain threshold. Inspired by the traditional script analysis studies, we conduct a human annotation study in the online customer review domain. Instead of asking human subjects to write down the elements important for evaluating review helpfulness by recalling from memory, we instruct participants to highlight the words and phrases that they consider to be important for making a review helpful. These review words are the ones that resonate in the readers' minds. Our proposed method for identifying script elements resembles the "recognition" method in memory research (Weiten, 2008). We also ask the participants to assess to what extent a review is helpful to a reader who is potentially interested in the object or event. We use a participant's review helpfulness score as a subjective measure of his/her perception of the review.

The common terms highlighted by the participants are indicators of the shared expectations, interests, and things looked for when determining whether a review is helpful. In a particular customer review domain (such as books or restaurants), the more a term is highlighted across a set of reviews, the more likely it is that the term is a salient element, reflecting what people expect to see in an online review for a particular domain. Therefore, review terms highlighted by multiple participants are the embodiment

of a shared cognitive script about the review subject. Script terms capture important aspects regarding an object/event (product or service reviewed), which are culturally shared and agreed upon. Hence, those reviews that properly address and elaborate on these salient script elements are more likely to be considered as useful. These explicitly elicited script terms are assembled into a script lexicon to represent domain-specific knowledge, which is used along with the assessed review helpfulness score, to construct script-based text regression models.

3.2. Conceptual models

In this section, we present the various models for predicting review helpfulness. The unit of study is an individual review. Each observation in the dataset is a unique review. In all the models we examine in this paper, the target variable is *review helpfulness*. Review helpfulness, as discussed before, is measured by the participants' answer to the helpfulness question. Figure 1 graphically depicts the development of the different conceptual models examined in this study.

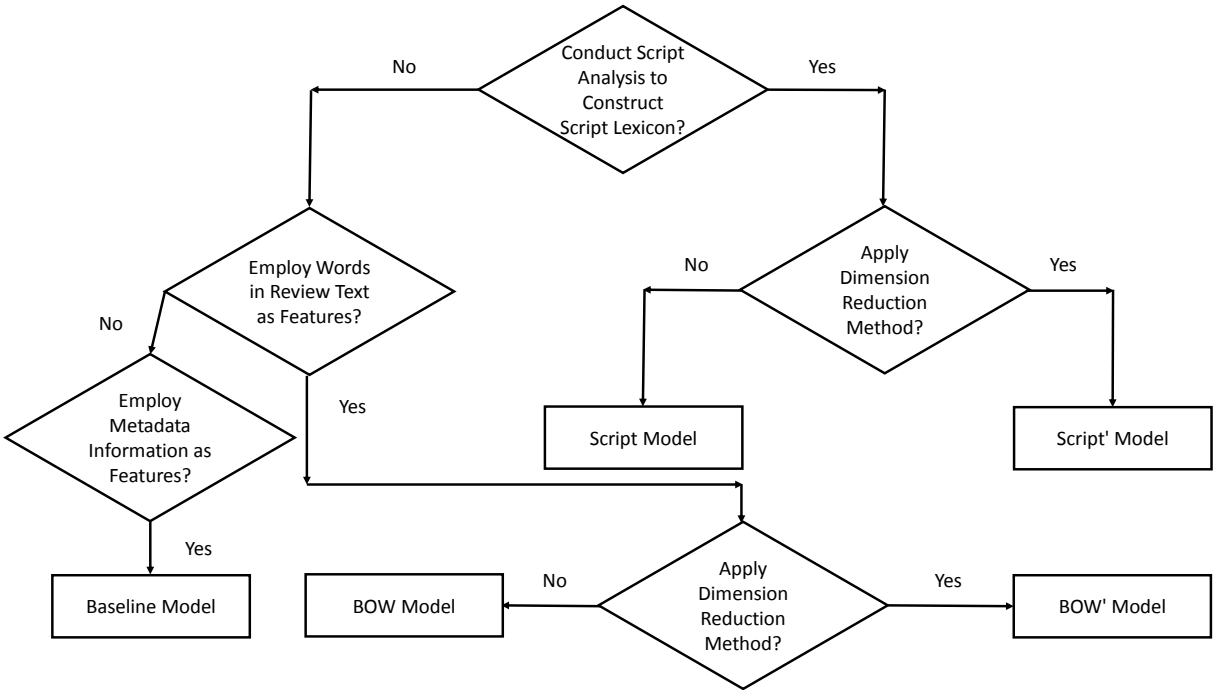

Fig. 1. Development of conceptual models.

3.2.1. Baseline models

In the literature on estimating review usefulness, various non-textual features have been examined. Non-textual features refer to the important metadata for online reviews, not the actual text (words) of the reviews. We develop a Baseline model with common constructs (metadata) reported in the state-of-the-art review helpfulness prediction models.

Review depth can be measured by review length or the number of words in a review. A longer review contains more product/service information and helps readers' decision process. Review length was found to be positively associated with review usefulness (Mudambi & Schuff, 2010). Review valence, measured by star rating (the number of stars a review author gives to the subject being reviewed), was also found to be an important variable influencing readers' perception of review usefulness (Mudambi & Schuff, 2010). Review extremity, computed as the difference between the current review's star rating and the average star rating of all reviews for the subject reviewed, was found to affect review helpfulness (Krishnamoorthy, 2015; Mudambi & Schuff, 2010).

Readability gauges how easy or difficult for a text to be comprehended by a reader. Readability was shown to be useful in estimating review usefulness (Ghose & Ipeirotis, 2011; Krishnamoorthy, 2015). Review age (how long has a review has been published), based on review publication date, was found to have an effect on review helpfulness (Krishnamoorthy, 2015; Liu, Cao, Lin, Huang, & Zhou, 2007). A review can express the author's positive or negative opinions. The number of positive and negative sentiment words was used to operationalize the construct of review text sentiment (Krishnamoorthy, 2015).

3.2.2. BOW-based models

Due to the representational complexities associated with text, the choice of features and feature selection techniques is important in text analysis systems (Abbasi & Chen, 2008). The standard BOW model keeps all the original features without applying any dimension reduction. The predictor variables are all the words that appear in a collection of reviews.

In general, text mining is characterized by a huge feature space and requires an effective and efficient feature selection procedure (Chou, Sinha, & Zhao, 2010; Ngo-Ye & Sinha, 2012). A dimension reduction technique could be used to narrow down the set of predictor variables. The reviews that we have collected for this study contain thousands of unique words (variables), which contain a lot of irrelevant, redundant, and noisy information. Hence, we employ feature selection to reduce dimensionality.

Since the original BOW model contains many noisy and irrelevant words, we apply a dimension reduction method to the BOW model to refine it and derive the dimension-reduced BOW model (BOW′, in short). BOW′ is a BOW-based model without knowledge of human concept scripts. It uses a feature selection technique to narrow down the set of predictor variables. We employ the Correlation-based Feature Selection (CFS) technique (see section 5.2.3.) to select a subset of review words that have high correlation with the target variable and low correlation within the set as features for the BOW′ model.

3.2.3. Script-based models

To examine the efficacy of script analysis for predicting review helpfulness, we construct two types of Script text regression models (using highlighted review words as predictors) and compare them with the BOW-based models. Although the predictor variables in both BOW-based models and script models are stemmed words, there are important differences regarding the selection criteria for the words. The crucial distinction is whether a concept script is involved in selecting the words. In the BOW models, no concept script is involved; they simply consider words that appear in the review text collection. In the script models, the words belong to the script lexicon. Therefore, the fundamental conceptual difference between these two types of models is that script models take advantage of the shared human knowledge, and concentrate on keywords from the concept script lexicon, while BOW models do not make use of this knowledge.

We now introduce the focus of this research – script-based models, which employ inputs from human concept scripts. The participants' highlighted phrases and words are assembled into a script

lexicon. These script words give us insights into the readers' mental judgment of which elements or terms are deemed useful in making a review helpful. The Script model keeps all the original highlighted words.

We further refine the Script model by constructing the dimension-reduced Script model (Script′, in short). Script′ is a script-based model using a lexicon derived from words highlighted by the participants. The difference between the Script′ model and the Script model is that Script′ has reduced dimensionality. We apply the CFS feature selection technique to select a subset of script words to form the reduced Script′ model. These script words have a high correlation with the target variable – review helpfulness – and low correlation among themselves.

Table 1 summarizes the three types of conceptual models of interest for this research.

Table 1 Conceptual framework of three text regression models for predicting online review helpfulness

Model Type	*Predictor Variables*	*Target Variable*
Baseline Model	review depth (length), review readability, review sentiment, review valence, review extremity, and review age	Review Helpfulness (scale 1-7)
BOW' Model	a subset of review words (tokenized unigrams) selected through applying CFS dimension reduction technique	Review Helpfulness (scale 1-7)
Script' Model	review words highlighted by human annotators, further reduced through applying CFS dimension reduction technique	Review Helpfulness (scale 1-7)

4. Research Questions

Based on script-theoretic arguments for analyzing the helpfulness of online reviews, we now propose a set of research questions and compare the performance of the three types of conceptual models presented in Table 1. In this research, we consider both the predictive accuracy of the models and their computational efficiency, similar to prior research (see, for example, (Chou, Sinha, & Zhao, 2010)).

4.1. Baseline versus Script′

While the metadata features of the Baseline model have been shown to be useful in predicting review helpfulness, they do not directly represent reviews' textual content at all. Besides, there is no human input in the Baseline model. Therefore, with the critical review textual information absent in the Baseline

model, its predictive accuracy of review helpfulness is expected to be limited. On the other hand, Script words, a subset of review words, do capture review textual information. Also, the participants considered the review words (script terms) they highlighted to contribute to review helpfulness. Hence, script-based models exploit the valuable shared domain knowledge that humans can offer. Equipped with this form of crucial human intelligence of reviews, Script models can be expected to estimate review helpfulness more accurately than the Baseline model. Moreover, the Script′ model has a more focused feature set, leading to an even more accurate prediction of review helpfulness. Hence, we pose the following research question:

RQ1a: *Is the Script′ model more accurate than the Baseline model in predicting review helpfulness?*

Computational efficiency is another important consideration for a model to be practical. We are also interested in comparing the computational efficiency of the Baseline model and the Script′ model. Hence, we pose the following research question:

RQ1b: *Is the Script′ model more computationally efficient than the Baseline model in predicting review helpfulness?*

4.2. BOW′ versus Script′

The features in BOW models are words appearing in reviews. While BOW models do represent review textual information, they contain many irrelevant and noisy words, and do not employ any human knowledge of the domain. On the other hand, the features in Script models are the highlighted review words, deemed useful by participants to make a review helpful. Thus, the script words are more likely to better represent the important elements that make a review useful than the features in BOW models. Overall, therefore, we expect Script models to be more accurate than BOW models. Furthermore, we applied a dimension reduction method to both BOW and Script models to generate the BOW′ and Script′ models. That further narrowed down the full Script feature set to the most salient elements for estimating review helpfulness. Because the Script′ models directly benefit from human domain knowledge, which the BOW′ models do not, we pose the following research question:

RQ2a: *Is the Script′ model more accurate than the BOW′ model in predicting review helpfulness?*

Also, because the Script model has much fewer features than the BOW model to start with and the same dimension reduction method is applied to both models, we expect that the dimensionality of the Script′ model to be smaller than that of the BOW′ model. Therefore, we pose the following research question:

RQ2b: *Is the Script′ model computationally more efficient than the BOW′ model in predicting review helpfulness?*

5. Methodology

We first describe the sources of data and the human annotation study we conducted with participants to create the script lexicon. Next, we report the text mining experiments that we carried out for comparing the script models with the baseline and the BOW models. Figure 2 outlines the proposed methodology for predicting the helpfulness of online customer reviews, which involves the following steps: review collection, review annotation to generate the script words, text preprocessing, index weighting, dimension reduction, model construction, and assessment of model performance.

We collected customer reviews from two influential review websites – Amazon.com and Yelp.com. Although both websites feature a wide variety of subjects being reviewed, we only selected the category of review subject that each website is most known for. From Amazon.com, we randomly chose nine books and retrieved all the reviews for those books. From Yelp.com, we randomly chose three restaurants and retrieved all the reviews for those restaurants. We have 1381 Amazon book reviews and 1219 Yelp restaurant reviews in the initial data pool.

5.1. Human annotation study

For the annotation study, we recruited a total of 135 undergraduate students enrolled in different classes at a large U.S. business school. Participation in this research activity was voluntary. The participants had the option to withdraw from the experiment at any time. Extra credit points were offered as an incentive to encourage participation. We assigned an identifier (ID) to each individual review in our data pool (1381 Amazon book reviews and 1219 Yelp restaurant reviews). Therefore, we have a total of 2600 (1381 +

1219) unique reviews. Because different classes had different amount of time available for the lab study, we arranged sets of eight reviews for some classes and sets of 12 reviews for the other classes. Each of 50 participants evaluated a set of eight reviews, with each set different from the other. Each of the remaining 85 participants evaluated a set of 12 reviews. Therefore, 135 participants evaluated a total of 1420 reviews. We randomly selected 8 or 12 reviews about one particular subject (a single book in the case of Amazon.com or a single restaurant in the case of Yelp.com). Then we assembled the 8 or 12 review texts into a single Microsoft Word document. The text of each individual review occupied a single page in the Word document and was labeled with a review ID. In this way, the set of reviews presented to a participant was all about the same review subject (a single book or a single restaurant). This arrangement mimics a real-world setting, where a reader visits a review website and reads a set of reviews about a single product/service. In the Word document, after each page of text of an individual review, we posed the following review helpfulness question: "Overall, to what extent is the review helpful to a reader who is potentially interested in buying the book/dining at the restaurant?" – on a 7-point Likert scale (where 1 means "none" and 7 means "a lot").

At the beginning of the annotation session, we briefly explained the research objective and provided instructions on the procedure. For each individual review, a participant performed the following tasks: 1) highlighted the phrases or words that he or she thought were useful in making the review helpful; 2) answered the review helpfulness question. Figure 3 shows an example of annotated review text.

Most participants completed the study diligently. However, some participants did not fully evaluate the reviews assigned to them. The helpfulness question was not answered for some reviews. Moreover, some participants did not highlight anything for all the reviews assigned to them. To ensure data quality, we removed those unusable reviews with a missing value in the helpfulness question, or those that did not have any highlights. The final valid sample size for Amazon book reviews is 757 and that for Yelp restaurant reviews is 570. We refer to the final Amazon dataset as A757 and to the final Yelp dataset as Y570.

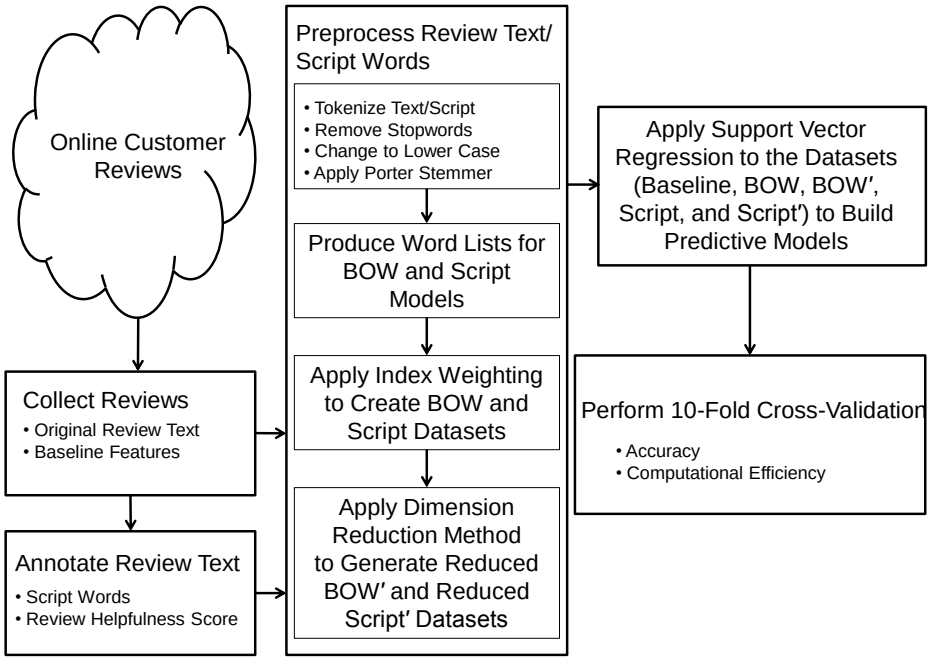

Fig. 2. Methodology for predicting online review helpfulness.

This is review number 101

Hey, I'm all about supporting the little guy but I feel that San Franciscans will hate on any business that is part of a National chain.

Sure the interior is a bit sterile here but it's nice enough. The sandwiches I've ordered have been delicious (albeit expensive) & you get a choice of a side including a slice of warm fresh bread. Service is quick, friendly and easy: an effortless ordering process. Beyond the sandwiches though, their food is hit and miss. I've had a couple unremarkable soups and an array of so-so baked goods and pastries. I expect a lot more at the prices they ask.

Fig. 3. An example of highlighted phrases for a review.

5.2. Text regression experiments

The main goal of this research is to explore the utility of script analysis for predicting online review helpfulness. In this empirical study, we perform text mining experiments to address the research questions, which involve comparing the Baseline, BOW′, and Script′ models (see Table 1).

5.2.1. Datasets for Baseline model

In the Baseline model, we include the following constructs: Review Depth (Length), Review Text Readability Scores, Review Text Sentiment, Review Valence (Star Rating), Review Extremity, and Review Age. The features employed in the Baseline model are widely reported in the literature on review helpfulness (Ghose & Ipeirotis, 2011; Krishnamoorthy, 2015; Mudambi & Schuff, 2010). We operationalize the constructs in the Baseline model with the following 14 predictor variables: Words, Characters, Paragraphs, Sentences, SentencesPerParagraph, WordsPerSentence, CharactersPerWord, FleschReadingEase, FleschKincaidGradeLevel, posemo, negemo, ElapsedDays, Rating, and ReviewExtremity. To obtain the value for the first nine variables (for constructs: Review Depth and Readability Scores), we wrote a VBA program to process review text and compute the values for these variables for each review. To gauge review text sentiment (positive emotion and negative emotion), we applied LIWC (Pennebaker, Francis, & Booth, 2001) to analyze review text to derive positive and negative sentiment value for each review. To obtain the value of the last two variables, we gathered review postdate and star rating for each review and performed calculations in Excel as shown in Table 2.

Table 2 Operationalization of predictor variables and target variable for baseline model

Constructs	**Variables Definition and Operation**
Review Helpfulness	The target variable is Review Helpfulness (scale 1-7)
Review Depth (Length)	Words = The number of words of a review Characters = The number of characters of a review Paragraphs = The number of paragraphs of a review Sentences = The number of sentences of a review
Review Readability	SentencesPerParagraph, WordsPerSentence, CharactersPerWord, FleschReadingEase, FleschKincaidGradeLevel
Review Sentiment	LIWC dictionary positive and negative dimensions (posemo and negemo)
Review Valence (Star Rating)	Rating = how many stars the review author assigned to the current review
Review Extremity	ReviewExtremity = Rating - AverageRating, difference between the current review rating (review valence) and average rating of all reviews for the book/restaurant reviewed
Review Age	ElapsedDays = DataCollectionDate – PostDate, how long the review has been posted online

Since the Baseline datasets do not employ individual review words as predictors, text preprocessing processes such as stemming and index weighting are not applicable. We constructed two Baseline datasets, one for Amazon and one for Yelp.

5.2.2. Datasets for BOW and Script models

Review content, represented by BOW models, has been found effective in estimating review helpfulness (Ngo-Ye & Sinha, 2012). Based on the script-theoretic arguments made in the previous section, script-based models would characterize review text content even better. In this section, we report the preprocessing procedures to generate the BOW and Script datasets. To generate the datasets for the BOW model, we perform the following procedure, as in (Ngo-Ye & Sinha, 2014), to extract the word lists from Amazon and Yelp review text collections. We tokenize the aggregated review text by removing all non-alphabetic characters. Then we apply the Standard English stopword filter to remove common stopwords, which do not carry significant meaning. Next, we change all the terms to lower case. We then apply the popular Porter Stemmer to reduce terms to their basic form. Finally, we obtain all the unique stemmed words that appear in a particular review text collection (Amazon or Yelp). We use these words as predictor variables in the BOW model.

To generate the datasets for the Script models, instead of starting from a collection of aggregated review text, we begin with the aggregated highlighted words collected from the annotation study (separately for Amazon and Yelp). Following the same preprocessing procedure as that of the BOW model, we obtain all the script words to be used as predictor variables in the Script model.

Using the above procedure, we derive the features of the BOW and Script models for Amazon and Yelp. With the predictor variables clearly specified, we next instantiate the datasets with realized values. In the datasets, each column corresponds to a unique stemmed word and each row corresponds to a unique review. To populate a text regression dataset, we make use of two types of index weighting schemes – Binary Occurrence and Term Occurrence (Sebastiani, 2002). We experiment with both index weighting schemes to assign a term weight to each feature in a review text. In the Binary Occurrence scheme, to represent the absence or presence of a term in a review text, we use binary value of 0 or 1. The number of times that a term appears in a review text is used as that term's weight in the Term Occurrence scheme. We refer to Binary Occurrence scheme as BinaOccu and Term Occurrence scheme as TermOccu.

5.2.3. Correlation-based feature selection (CFS)

The rudimentary rationale for CFS is that the preferred feature subset includes features that are prominently correlated with the target variable, yet have small correlations among the features themselves (Hall, 2000). The "merit" or worth of a subset of features for estimating the target variable can be defined in the following formulas (Hall, 2000). Presume that a feature subset *S* contains *n* features {f_1, f_2, f_3, …, f_n} and the target variable is *t*. The average feature-target correlation $\overline{r_{ft}}$ is defined as:

$$\overline{r_{ft}} = \frac{\sum_{i=1}^{n} correlation(f_i,t)}{n}.$$

The average feature-feature inter-correlation $\overline{r_{ff}}$ is defined as:

$$\overline{r_{ff}} = \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} correlation(f_i,f_j)}{n*(n-1)}$$

The merit of S, a subset of features, can be defined as:

$$Merit_S = \frac{n * \overline{r_{ft}}}{\sqrt[2]{n + n*(n-1)*\overline{r_{ff}}}}$$

After calculating the correlation matrix from the data, the best-first search strategy is usually employed to search in the subset space (Witten, Frank, & Hall, 2011). We denote this technique of CFS subset evaluation with best-first search as CfsBF. This greedy best-first approach for CFS is quite computationally efficient (Chou, Sinha, & Zhao, 2010). CfsBF has been shown as a competitive feature selection technique for the regression problem (Hall, 2000; Hall & Holmes, 2003). CfsBF has also been successfully applied as a dimension reduction method in text regression (Ngo-Ye & Sinha, 2012; 2014).

We employ CfsBF as a feature selection technique to study the effects of dimension reduction. Because we have coded all the original features as numeric variables, we can use the conventional linear Pearson's correlation formula in CFS. We apply the CfsBF dimension reduction technique to the corresponding original datasets (BOW and Script) to generate BOW′ and Script′ datasets.

5.2.4. Summary of instantiated datasets

While the central goal of the text regression experiment is to compare the Script′ model with the BOW′ model and the Baseline model, we need to construct various datasets of direct interest (Script′, BOW′, and Baseline) and supporting datasets (Script and BOW), upon which Script′ and BOW′ are built. Next, we summarize the instantiated text regression datasets. Overall, we have two types of datasets. The first type is BOW-based and Script-Based, which makes use of individual review words as variables. A total of 16 datasets belong to the first type. The second type is Baseline, which does not literally employ words from review text as variables. A total of two datasets belong to the second type, one for Amazon and one for Yelp. In Table 3, we present the 18 instantiated datasets (16 BOW-based and Script-Based datasets plus two Baseline datasets) with review helpfulness as the target variable.

For BOW-based and Script-Based datasets, first, we have four sub-types, which are: BOW, Script, BOW′, and Script′. Second, we have two sources of review text collection (Amazon and Yelp). Third, we try two types of index weighting schemes (Binary Occurrence and Term Occurrence). Therefore, we have constructed 16 (4 X 2 X 2) BOW-based and Script-Based datasets based on the combinations mentioned above. To produce these 16 instantiated datasets, we employ the RapidMiner data mining tool (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006) to perform the text preprocessing and index weighting tasks.

Moreover, we construct two Baseline datasets by using Excel formula, LIWC package, and VBA custom program to calculate the values of the variables for these two datasets. We choose Weka ARFF file format (Witten, Frank, & Hall, 2011) as the output format for all the 18 datasets.

We next report the number of variables, including the target variable (review helpfulness) for the 18 instantiated datasets. Table 4 shows that BOW′ datasets have far fewer variables than BOW datasets, because CfsBF reduces the dimensionality significantly. Comparing Script datasets with BOW datasets, the number of unique highlighted review words by the human annotators is about one third of all the unique review words present in the review text collection. In other words, the Script analysis method

reduces the dimensionality of BOW to about one third. We also find that applying CfsBF to Script datasets further reduces the dimensionality to an even smaller number. In fact, the dimensionality of Script′ is less than one third of BOW′'s dimensionality. The size of script lexicon ranges from 164 (Amazon BinaOccu dataset excluding target variable) to 226 (Amazon/Yelp TermOccu datasets excluding target variable). These amount of script variables are used in text regression models. Note that the Baseline datasets have the fewest number of variables (14 excluding target variable).

Table 3 18 instantiated datasets with review helpfulness as target variable

Data Source	**Index Weighting**	**BOW**	**BOW′**
Amazon	BinaOccu	A757BinaOccu	A757BinaOccuCfsBF
	TermOccu	A757TermOccu	A757TermOccuCfsBF
Yelp	BinaOccu	Y570BinaOccu	Y570BinaOccuCfsBF
	TermOccu	Y570TermOccu	Y570TermOccuCfsBF
Data Source	**Index Weighting**	**Script**	**Script′**
Amazon	BinaOccu	A757BinaOccuS	A757BinaOccuSCfsBF
	TermOccu	A757TermOccuS	A757TermOccuSCfsBF
Yelp	BinaOccu	Y570BinaOccuS	Y570BinaOccuSCfsBF
	TermOccu	Y570TermOccuS	Y570TermOccuSCfsBF
Data Source	**Baseline**		
Amazon	A757Baseline		
Yelp	Y570Baseline		

Table 4 Number of variables for the text regression datasets (excluding target variable)

Data Source	*Index Weighting*	*BOW*	*BOW′*	*Script*	*Script′*	*Baseline*
Amazon	BinaOccu	4394	529	1465	164	14
	TermOccu	4394	736	1465	226	
Yelp	BinaOccu	4595	664	1469	206	14
	TermOccu	4595	742	1469	226	

5.2.5. Text regression algorithm and experiment configurations

We next describe the design of the text mining experiments to predict review helpfulness. Support vector regression (SVR), a margin-based learner, has been demonstrated to be effective in predicting review helpfulness in previous studies (Kim, Pantel, Chklovski, & Pennacchiotti, 2006; Zhang, 2008; Zhang & Varadarajan, 2006). Therefore, we adopt SVR as the text regression algorithm for this study. We employ

the state-of-the-art text regression algorithm LibSVM (Chang & Lin, 2001; EL-Manzalawy & Honavar, 2005). We experiment with two types of SVRs – nu-SVR and epsilon-SVR. We use the radial basis function (RBF), which is the default kernel choice for both LibSVM nu-SVR and LibSVM epsilon-SVR. Hereafter, we refer to the instantiation of the algorithms as LibSVMnuRBF and LibSVMepRBF. Both types of SVR generate similar results. In the interest of space, we only report the results for LibSVMnuRBF.

We apply the text regression algorithms LibSVMnuRBF and LibSVMepRBF to the 18 datasets to generate and evaluate models. We use 10-fold cross-validation for reliable estimation (Witten, Frank, & Hall, 2011). We repeat the cross-validation for 10 experimental runs and use the average performance across the 10 runs as the final estimate for a particular model. Therefore, we have 18 datasets X 2 regression algorithms X 10 runs X 10 fold cross-validation = 3600 observations of performance. We aggregate the 100 observations of performance for each dataset and regression algorithm. This overall average across 10 runs and 10-fold cross-validation is reported as the text regression performance for each model in this paper.

5.2.6. Regression performance measures

Predictive analytics comprise not only empirical methods that produce data predictions but also methods for evaluating predictive power (Shmueli & Koppius, 2011); for a model to be practically useful, it needs to predict well. Before reporting the details of the model comparisons, we first describe the measures of regression performance. We evaluate different conceptual models based on four error-based regression performance measures, as well as two computing time-based measures. Error-based measures essentially gauge, on average, the residual or the difference between the actual observation and the predicted target variable. We employ four kinds of such error-based measures – *Root Mean Squared Error, Root Relative Squared Error, Mean Absolute Error, and Relative Absolute Error*.

Root Mean Squared Error (RMSE) uses the square root of the average squared loss to measure regression error, i.e.,

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

The default rule, ZeroR, is a trivial algorithm which always predicts the target class value of a test case as the average target class value in the training set. The RMSE of ZeroR model is defined as:

$$RMSE_{ZeroR} = \sqrt[2]{\frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}{N}}$$

Root Relative Squared Error (RRSE) takes the regression algorithm's RMSE and divides it by ZeroR's RMSE (Gama & Brazdil, 1995). Therefore, Root Relative Squared Error is a relative measure, comparing how better off of a learner against the default rule.

$$RRSE = \sqrt[2]{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}}$$

Instead of using squared loss function as in Root Mean Squared Error, Mean Absolute Error (MAE) uses the absolute differences between actual observation and predicted target variable.

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N}$$

The MAE of the ZeroR model is defined as:

$$MAE_{ZeroR} = \frac{\sum_{i=1}^{N}|y_i - \bar{y}_i|}{N}$$

Relative Absolute Error (RAE) takes the regression algorithm's MAE and divides it by the ZeroR's MAE. Thus it is also a relative measure as Root Relative Squared Error.

$$RAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|y_i - \bar{y}_i|}$$

We also use two measures for computational efficiency: *user CPU time training* and *user CPU time testing*, available in Weka. These two measures are the exact CPU times (in seconds) for training and testing the text regression models. For all the four error-based measures, as well as the two running time-based measures, the smaller the realized value, the better is the performance of the regression model. Next, we compare the predictive performance of the three conceptual models.

6. Results of text mining experiments

For two models to be comparable, we control the following three factors. First, we have two sources of review text collection (Amazon and Yelp). Second, we use two types of index weighting schemes (Binary Occurrence and Term Occurrence). Third, we consider two regression algorithms: LibSVMnuRBF and LibSVMepRBF. The combination of these control factors creates unique scenarios for model comparison. Therefore, we have a total of 2 X 2 X 2 = 8 scenarios to examine for each pair of models. In each scenario, we compare two models across six types of regression performance measures. To tackle the research questions, we conduct pairwise model comparisons and highlight the cells of the better performing models. The values presented in the following tables are the average regression performance measures over 10 experimental runs and 10-fold cross-validation. Moreover, we perform paired *t*-tests and report whether the differences are statistically significant.

6.1. Baseline versus Script′

In research question RQ1, we are interested in examining whether using a set of CfsBF-reduced highlighted script words as features in text regression would perform better than the Baseline model for predicting review helpfulness. In this section, we compare the regression performance of the Baseline and the Script′ model. For the Baseline model, because it does not make use of individual words in a review, the index weighting schemes have no impact. Therefore, the performance result of Baseline is the same across Binary Occurrence and Term Occurrence. First, we present the results of Amazon dataset with LibSVMnuRBF (see Table 5). Every T-value reported in the table is high because the mean difference in performance between the two models, with respect to all the performance measures, is much higher than the standard error of the mean difference. In all the 12 cases of Amazon dataset with LibSVMnuRBF, Script′ models always significantly outperform the Baseline model ($p < 0.001$).

Next, we present the results for Yelp dataset with LibSVMnuRBF. Table 6 shows that in all the 12 cases of Yelp dataset with LibSVMnuRBF, Script′ models always perform significantly better than the Baseline model. We also find that when using LibSVMepRBF, for both Amazon and Yelp datasets,

Script′ models always perform better than Baseline models. The performance differences between these two models are highly significant ($p < 0.001$) in all 24 cases of LibSVMepRBF. The results, taken as a whole, indicate that Script′ models are considerably more accurate and efficient than Baseline models in predicting review helpfulness. Therefore, research questions RQ1a and RQ1b are strongly supported.

Table 5 Comparison of Baseline models with Script′ models for predicting review helpfulness, LibSVMnuRBF, Amazon

Performance Measures	*Index Weighting*	*Baseline*	*Script'*	*T-Value*	*P-Value*
MAE	BinaOccu	1.24674	**1.15864**	14.802	0.000**
	TermOccu	1.24674	**1.09117**	21.670	0.000**
RMSE	BinaOccu	1.52101	**1.42702**	15.135	0.000**
	TermOccu	1.52101	**1.34695**	24.865	0.000**
RAE	BinaOccu	98.47431	**91.54461**	34.274	0.000**
	TermOccu	98.47431	**86.25498**	29.687	0.000**
RRSE	BinaOccu	99.93061	**93.71935**	41.266	0.000**
	TermOccu	99.93061	**88.49336**	35.441	0.000**
Training Time (in seconds)	BinaOccu	0.41777	**0.11684**	86.303	0.000**
	TermOccu	0.41777	**0.14695**	73.317	0.000**
Testing Time (in seconds)	BinaOccu	0.03463	**0.00905**	17.293	0.000**
	TermOccu	0.03463	**0.01108**	16.109	0.000**

Note: ** denotes a significant level of $p < 0.001$

Table 6 Comparison of Baseline models with Script′ models for predicting review helpfulness, LibSVMnuRBF, Yelp

Performance Measures	*Index Weighting*	*Baseline*	*Script'*	*T-Value*	*P-Value*
MAE	BinaOccu	1.35260	**1.31523**	7.107	0.000**
	TermOccu	1.35260	**1.28617**	11.464	0.000**
RMSE	BinaOccu	1.62426	**1.57257**	9.390	0.000**
	TermOccu	1.62426	**1.54453**	13.895	0.000**
RAE	BinaOccu	99.16292	**96.44980**	19.495	0.000**
	TermOccu	99.16292	**94.31991**	23.007	0.000**
RRSE	BinaOccu	100.14163	**96.91301**	25.091	0.000**
	TermOccu	100.14163	**95.21034**	29.889	0.000**
Training Time (in seconds)	BinaOccu	0.21809	**0.07223**	46.706	0.000**
	TermOccu	0.21809	**0.07597**	44.135	0.000**
Testing Time (in seconds)	BinaOccu	0.01685	**0.00577**	9.229	0.000**

	TermOccu	0.01685	**0.00640**	8.326	0.000**

Note: ** denotes a significant level of $p < 0.001$

6.2. BOW′ versus Script′

In research question RQ2, we are interested in examining whether using a set of CfsBF-reduced highlighted script words as features in text regression would perform better than CfsBF-reduced BOW models for predicting review helpfulness. In this section, we compare the regression performance of BOW′ and Script′ models. First, we present the results of Amazon dataset with LibSVMnuRBF.

Table 7 shows that in all the 12 cases of Amazon dataset with LibSVMnuRBF, Script′ models always significantly outperform BOW′ models ($p < 0.001$). Next, we present the results for Yelp dataset with LibSVMnuRBF.

Table 8 shows that in all the 12 cases of Yelp dataset with LibSVMnuRBF, Script′ models always perform significantly better than BOW′ models. We also find that when using LibSVMepRBF, for both Amazon and Yelp datasets, Script′ models always perform better than BOW′ models. The performance differences between these two models are highly significant ($p < 0.001$) in 23 out of the 24 cases of LibSVMepRBF; only in one case, the difference is not significant ($p = 0.090$). The results, taken as a whole, indicate that Script′ models are considerably more accurate and efficient than BOW′ models in predicting review helpfulness. Therefore, research questions RQ2a and RQ2b are also strongly supported.

Table 7 Comparison of BOW′ models with Script′ models for predicting review helpfulness, LibSVMnuRBF, Amazon

Performance Measures	*Index Weighting*	*BOW′*	*Script′*	*T-Value*	*P-Value*
MAE	BinaOccu	1.19793	**1.15864**	20.700	0.000**
	TermOccu	1.13546	**1.09117**	23.291	0.000**
RMSE	BinaOccu	1.51011	**1.42702**	52.697	0.000**
	TermOccu	1.39430	**1.34695**	25.737	0.000**
RAE	BinaOccu	94.62047	**91.54461**	21.294	0.000**
	TermOccu	89.73034	**86.25498**	24.381	0.000**
RRSE	BinaOccu	99.17759	**93.71935**	54.523	0.000**
	TermOccu	91.57516	**88.49336**	27.788	0.000**
Training Time (in seconds)	BinaOccu	0.24352	**0.11684**	26.893	0.000**
	TermOccu	0.30467	**0.14695**	30.416	0.000**

Testing Time (in seconds)	BinaOccu	0.01888	**0.00905**	7.757	0.000**
	TermOccu	0.02356	**0.01108**	8.299	0.000**

Note: ** denotes a significant level of $p < 0.001$

Table 8 Comparison of BOW′ models with Script′ models for predicting review helpfulness, LibSVMnuRBF, Yelp

Performance Measures	*Index Weighting*	*BOW′*	*Script′*	*T-Value*	*P-Value*
MAE	BinaOccu	1.35195	**1.31523**	26.492	0.000**
	TermOccu	1.33998	**1.28617**	26.395	0.000**
RMSE	BinaOccu	1.61461	**1.57257**	20.421	0.000**
	TermOccu	1.60065	**1.54453**	18.687	0.000**
RAE	BinaOccu	99.13940	**96.44980**	27.835	0.000**
	TermOccu	98.26069	**94.31991**	27.216	0.000**
RRSE	BinaOccu	99.47876	**96.91301**	21.706	0.000**
	TermOccu	98.62712	**95.21034**	19.849	0.000**
Training Time (in seconds)	BinaOccu	0.17238	**0.07223**	29.707	0.000**
	TermOccu	0.14945	**0.07597**	19.483	0.000**
Testing Time (in seconds)	BinaOccu	0.01388	**0.00577**	6.323	0.000**
	TermOccu	0.01108	**0.00640**	3.238	0.002*

Note: ** denotes a significant level of $p < 0.001$, and * denotes a significant level of $p < 0.01$.

We also compare BOW and Script models to examine whether using all highlighted script words as features in text regression would outperform using all words that appear in a collection of review text for predicting review helpfulness. We find that in all the 12 (two index weighting schemes X six types of regression performance measures) cases of Amazon dataset with LibSVMnuRBF, Script models always significantly outperform BOW models ($p < 0.001$). Similarly, we find that in all the 12 cases of Yelp dataset with LibSVMnuRBF, Script models also always perform significantly better than BOW models ($p < 0.001$). Likewise, we find that when using LibSVMepRBF, for all the 24 cases (12 cases for Amazon and 12 cases for Yelp), Script models always significantly outperform BOW models as well. The results strongly indicate that Script models are significantly more accurate and efficient than BOW models, demonstrating the efficacy of script analysis.

Finally, we compare Script and Script′ models to examine whether dimension reduction helps in improving the regression performance. We find that in all the 12 cases of Amazon dataset with LibSVMnuRBF, Script′ models always significantly outperform Script models ($p < 0.001$). Similarly, we

find that in all the 12 cases of Yelp dataset with LibSVMnuRBF, Script′ models always perform significantly better than Script models ($p < 0.001$) as well. Moreover, we find that when using LibSVMepRBF, for both Amazon and Yelp datasets, Script′ models always significantly outperform Script models ($p < 0.001$). Taken together, dimension-reduced Script′ models always perform significantly better than the corresponding original Script models in predicting review helpfulness. Therefore, employing the CfsBF method, in conjunction with script analysis, further enhances text regression performance.

In this section, we empirically compared the Baseline, BOW′, and Script′ models. The results consistently support both sets of research questions. Overall, Script′ models outperform both Baseline and BOW′ models. The results also demonstrate that Script′ models are always the best among the different conceptual models. The results of the text mining experiments show the efficacy of the reduced Script′ model. Scripts, especially the CfsBF-selected script words, capture important human expert knowledge. We can therefore conclude that a select group of script words are key to capturing review helpfulness.

6.3. Feature selection time

In the previous section, we evaluated the three conceptual models based on four error-based measures and two runtime-based measures. Another relevant measure – *feature selection time* – should also be considered, when appropriate. Among the models, additional feature selection time is required to derive the two reduced models (BOW′ and Script′). The CfsBF feature selection process is computationally intensive for the text regression problem and, therefore, it is important to track the times involved. Table 9 presents the times taken for the feature selection tasks (the unit of time is a minute). We highlight the better performing model in bold.

Table 9 Feature selection time for generating BOW′ and Script′ datasets (in minutes)

Data Source	*Index Weighting*	*BOW′*	*Script′*
Amazon	BinaOccu	324	**5**
	TermOccu	587	**11**

Yelp	BinaOccu	565	**8**
	TermOccu	721	**10**

Table 9 shows that the time for feature selection ranges from 5 minutes to 721 minutes. The feature selection processing times to generate the Script′ models are significantly shorter than that needed to generate the BOW′ models. Thus, the Script analysis has the additional benefit of reducing feature selection time substantially. The length of time required for feature selection should be taken into consideration when deploying a model in real-world settings where instantaneous processing time is desired. Hence, when interpreting the processing time of these two feature selection-based models (BOW′ and Script′), we should also include the extra time needed for the feature selection step. We find that the Script′ model becomes even more attractive when considering overall model processing time.

7. Discussion

We compared the Baseline, BOW′, and Script′ models and found that both sets of research questions were answered in the affirmative. Next, we elaborate on the results from the text mining experiments of Script′ versus Baseline and BOW′ models.

Our study demonstrates that the script models are superior to the Baseline and BOW models for predicting review helpfulness. With a much smaller set of script terms, the Script models are more accurate than the BOW models, which use all the words that appear in the review text collection. The implication is that not all words are equal in making a review helpful. The script words are more essential to human conception of a review subject. Cognitive scripts, as a form of human knowledge of a domain, help to filter out irrelevant noise presented in the review text. However, the Baseline and BOW models do not have access to such domain-specific knowledge. By incorporating the human knowledge of the review domain, Script models are able to predict review helpfulness more accurately and efficiently than BOW models.

Dimension-reduced Script′ models outperform both Baseline and dimension-reduced BOW′ models. Both Baseline and BOW′ models do not employ any human knowledge. The dimensionality reduction in BOW′ is achieved via purely technical means. It represents a data-driven induction approach.

On the other hand, Script′ models represent distilled human knowledge of the review domain obtained through the combination of script analysis and dimension reduction. Empirical evidence from our study shows that the distilled script knowledge approach is superior to the data-driven induction approach. This is a crucial discovery in that it highlights the value of human expert knowledge, cognitive script in this case. The business implication is that, when possible, human domain knowledge should be sought and incorporated into a model (Sinha & Zhao, 2008). The empirical findings demonstrate the effectiveness and power of the approach of combining script analysis with dimension reduction. It produces better results than pure text mining models alone can achieve. Thus, incorporating human expert knowledge can considerably enhance the performance of the state-of-art text mining models.

Moreover, dimension-reduced Script′ models further improve the regression performance of Script models. Script′ models are always more accurate than Script models. The implication is that a smaller subset of select script words is more salient for abstracting the review domain; it represents distilled human expert knowledge. This distilled human expert knowledge is the key to the success of Script′ models.

We can interpret CfsBF as a way of using a pure machine learning method to reduce dimensionality to improve text regression performance. On the other hand, Script analysis uses human expert knowledge to reduce dimensionality. Both approaches improve model performance. Moreover, the most accurate models are always the Script′ models. Therefore, combining script analysis and CfsBF appears to be the most attractive approach in terms of prediction accuracy.

The results of our study show that Script′ models have much smaller training and testing time compared to both Baseline and BOW′ models. A plausible reason might be that the predictors in Script' model are more relevant for predicting the target variable (review helpfulness) than those in the Baseline and BOW′ models. Moreover, the Script′ models further lower the training and testing times compared to the original Script models. In the event of real-time online analysis of customer reviews, computational efficiency will be a key factor in choosing models. Script-based models are much better with respect to

computational efficiency. Furthermore, the feature selection task consumes a substantial amount of computational resources and time. The Script′ models have a clear advantage over the BOW′ models in terms of feature selection time. Taken together, based on all the results on computing time and feature selection time, the Script′ models appear to be a much more attractive option than the Baseline and BOW′ models.

Moreover, script-based models produce some human interpretable and directly usable knowledge. It is relatively easy for management to understand the script analysis approach. Both the intuition of human expert knowledge and the support from psychology research of cognitive scripts make the results of the script model more interpretable and understandable. We need to point out that the easier it is for management to understand and interpret the text mining model, the more likely it is that the model will be valued and actually used. Script-based models have the strengths of having a foundation in psychology theory, higher effectiveness and efficiency as confirmed by empirical experiments, and ease of interpretation and understanding.

8. Conclusion and future directions

This paper contributes to the literature in several ways. First, the proposed script-based model is built upon a solid theoretical foundation. Compared with existing studies of cognitive script theory, our proposed script analysis approach makes several distinctive contributions: 1) None of the prior studies of cognitive script theory have examined the efficacy of using script analysis for predicting review helpfulness. Our research is the first study making the conceptual link between cognitive script theory and the application of estimating the helpfulness of online customer reviews. Our study goes beyond baseline features (metadata) and text features (bag-of-words) by incorporating higher level knowledge constructs – cognitive scripts – into a model to predict the helpfulness of online reviews. 2) The previous applications of script theory in business research in general, and Leigh and Rethans's (1984) study in particular, use a technique that resembles the "recall" method in memory research to elicit script items. Our research contributes to the application of cognitive script theory by pioneering a technique that resembles the "recognize" method in memory research. Drawing from the principle of script theory, we conduct an

annotation-based script analysis by explicitly eliciting script terms for helpful online reviews in a domain. These script terms are used to construct a script lexicon, which represents domain-specific knowledge. 3) While past script analysis studies count the frequency of appearances of script terms and report simple statistics, our research goes further by employing annotated script terms as new features in sophisticated text regression models and conducting advanced statistical analysis.

Second, with respect to methodology, we pioneer the application of the annotation method to script analysis in the domain of studying the helpfulness of online reviews. To the best of our knowledge, this is the first empirical study that directly extracts script elements for online customer reviews from human annotators. This research extends the traditional script analysis methodology by introducing an original annotation procedure to elicit script terms for helpful online reviews. Instead of asking participants to directly write down script terms from their mind, participants were presented with review text and asked to highlight useful phrases that resonated with their mental script. The highlighted review phrases by the participants provide useful first-hand information of the lexicon of online review scripts. This creative design of annotation-based script analysis has many potential applications in e-commerce and online social media research.

Third, we proposed and empirically validated a text regression model – Script′ – which combines script analysis and dimension reduction, for predicting online review helpfulness. Using an empirical annotation study and a set of text mining experiments, we demonstrated the superiority of script analysis over state-of-the-art review helpfulness prediction methods. The results of the text mining experiments show that cognitive script-based models, which involve the higher level mental constructs, predict the helpfulness of online customer reviews significantly better than the Baseline and the BOW′ models. Moreover, the results show that this concept scripts-enriched text regression model – Script′ – always performs the best among the models considered in this study. Therefore, a small subset of script words chosen by feature selection is needed to effectively predict review helpfulness. We attribute the success of the script-based text mining models to the fact that cognitive scripts capture essential human knowledge

of a specific review domain. The human expert knowledge enables script models to concentrate on the most salient aspects of human evaluation of review helpfulness.

The practical implication of this research is that the proposed script theory-based approach is a very useful alternative for building a review helpfulness prediction model. A domain specific script lexicon can be established by having just a few hundred reviews in a domain annotated by human subjects. After incurring this one-time cost, a script lexicon can be applied and reused for estimating the helpfulness of new reviews. Because the script lexicon, a collective mental representation of certain events or objects, is relatively stable, the script analysis approach is certainly feasible for practical use. This study empirically demonstrates the relative effectiveness and efficiency of the proposed scripts-enriched model for predicting review helpfulness, compared to the Baseline model and the BOW′ model. Review websites should adopt the Script′ model, rather than the Baseline model or the BOW′ model, to identify those reviews believed to be very helpful. Moreover, the "most helpful reviews" identified by the Script′ model can be presented prominently to end users for improving their information seeking experience and enhancing the review website's usefulness.

For review websites, product manufacturers, and service providers, our proposed script analysis model will help filter business intelligence by identifying the most helpful reviews. The script model will also provide some understandable and interpretable insights on consumer perception of the subject reviewed. Besides, it is relatively easy to communicate and explain the script analysis model to management. Moreover, our proposed model can help management to build a domain-specific lexicon with extracted salient terms from review text. The key concepts from the lexicon can be featured in marketing campaigns for the reviewed products or services, because they are considered relevant and important by consumers.

The script-based models are computationally more efficient than the Baseline and the BOW′ models. Script-based models require much less training, testing, and feature selection time. This makes

the script model a very attractive choice for identifying helpfulness reviews in real-world situations where processing time is a critical constraint.

The empirical results of this study are based on review collections from Amazon and Yelp. The findings are consistent across both datasets. We acknowledge that the script lexicon constructed in this study is specific for the book and restaurant domains, though for a particular domain, the script lexicon is relatively stable. However, to generalize the research conclusions, more studies are needed. Potential extensions of this work may include testing the proposed models on collections of review data from other domains such as electronic products, hotels, etc., and from other review websites. Also, we could examine if there are differences in the effectiveness of script-based models on helpfulness of online reviews for search and experience products (Baek, Ahn, & Choi, 2013; Mudambi & Schuff, 2010).

In this research, we applied script theory for the text regression problem. Another potential extension is to use script theory for the text summarization task. Script, by its very nature, is an outline or concise abstraction of an activity or an event. Therefore, theoretically, a cognitive script itself can be regarded as a summarization of a text. Using highlighted phrases in the online review text, we can identify the underlying cognitive scripts for a particular review domain. From the scripts, we can derive a more meaningful text summarization. Script-based summarization has the advantage of having a solid theoretical foundation, ease of interpretation, understanding, and communication. We plan to explore the potential of script-based summarization and expect new theoretical contributions to our knowledge of text summarization. Moreover, script-based summarization will also have practical applications, such as better summarization of online review text for consumers and businesses.

Several online retailers provide recommender systems to help consumers make purchase decisions. Baum and Spann (2014) have experimentally analyzed the interplay between recommendations provided by recommender systems and those of online reviews written by previous customers. In a similar vein, our study could be extended to examine if there are interactions between recommendations of recommender systems and those of script-based models.

Reviewer identity disclosure positively affects online community members' evaluation of the helpfulness of online reviews (Forman, Ghose, & Wiesenfeld, 2008). In a future study, we plan to evaluate a new model that combines script analysis and reviewer identity, which may produce better results.

In summary, our study shows that the script theory approach – more specifically, using annotated script terms as new features in text regression models – is very effective and efficient for predicting online review helpfulness. The theoretical implication is that script as a knowledge representation structure for a specific domain is instrumental in people's perception of the usefulness of a text. Another implication is that we can adapt the traditional script analysis with the annotation method to construct a script lexicon that represents human expert knowledge in a particular domain. We plan to further expand our knowledge of human evaluation of online social media text in different domains with the help of cognitive script theory.

References

Abbasi, A., & Chen, H. (2008, December). CyberGate: A System and Design for Text Analysis of Computer Mediated Communications. *MIS Quarterly, 32*(4), 811-837.

Abelson, R. P. (1976). Script Processing in Attitude Formation and Decision Making. In J. D. Carroll, & J. Payne (Eds.), *Cognition and Social Behavior* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Ainscough, T. L., DeCarlo, T. E., & Leigh, T. W. (1996). Building Expert Systems from the Selling Scripts of Multiple Experts. *The Journal of Services Marketing, 10*(4), 23-40.

Baek, H., Ahn, J., & Choi, Y. (2013). Helpfulness of online consumer reviews: readers' objectives and review cues. *International Journal of Electronic Commerce, 17*(2), 99-126. doi:10.2753/JEC1086-4415170204

Baek, H., Lee, S., Oh, S., & Ahn, J. (2015, October). Normative social influence and online review helpfulness: polynomial modeling and response surface analysis. *Journal of E-commerce Research, 16*(4), 290-306. Retrieved from http://www.jecr.org/node/476

Barley, S. R. (1986). Technology as an Occasion for Structuring: Evidence from Observations of CT Scanners and the Social Order of Radiology Departments. *Administrative Science Quarterly, 31*(1), 78-108.

Barley, S. R. (1990). Images of Imaging: Notes on Doing Longitudinal Field Work. *Organization Science, 1*(3), 220-247.

Baum, D., & Spann, M. (2014). The interplay between online consumer reviews and recommender systems: an experimental analysis. *International Journal of Electronic Commerce, 19*(1), 129-161.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in Memory for Text. *Cognitive Psychology, 11*, 177-220.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring Determinants of Voting for the "Helpfulness" of Online User Reviews: A Text Mining Approach. *Decision Support Systems, 50*(2), 511-521.

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM : a library for support vector machines.* Retrieved May 13, 2011, from http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen, H., & Zimbra, D. (2010). AI and Opinion Mining. *IEEE Intelligent Systems, 25*(3), 74-76.

Chen, J., Xu, H., & Whinston, A. B. (2011, Fall). Moderated online communities and quality of user-generated content. *Journal of Management Information Systems, 28*(2), 237-268. doi:10.2753/MIS0742-1222280209

Chou, C.-H., Sinha, A. P., & Zhao, H. (2010, September). A Hybrid Attribute Selection Approach for Text Classification. *Journal of the Association for Information Systems, 11*(9), 491-518.

Clemons, E. K., Gao, G., & Hitt, L. M. (2006, Fall). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems, 23*(2), 149-171. doi:10.2753/MIS0742-1222230207

EL-Manzalawy, Y., & Honavar, V. (2005). *WLSVM : Integrating LibSVM into Weka Environment.* Retrieved May 13, 2011, from http://www.cs.iastate.edu/~yasser/wlsvm/

Erasmus, A. C., Boshoff, E., & Rousseau, G. (2002). The Potential of Using Script Theory in Consumer Behavior Research. *Journal of Family Ecology and Consumer Sciences, 30*, 19.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008, September). Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Information Systems Research, 19*(3), 291-313.

Gama, J., & Brazdil, P. (1995). Characterization of Classification Algorithms. *Proceedings of EPIA 95, Progress in Artificial Intelligence; 7th Portuguese Conference on Artificial Intelligence. LNAI Vol.990*, pp. 189-200. Funchal, Madeira Island, Portugal: Springer-Verlag.

Ghose, A., & Ipeirotis, P. G. (2011, October). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498-1512. doi:10.1109/TKDE.2010.188

Hall, M. A. (2000). Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, (pp. 359–366).

Hall, M. A., & Holmes, G. (2003, May/June). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering, 15*(3), 1-16.

Huang, G.-H., & Korfiatis, N. (2015). Trying before buying: the moderating role of online reviews in trial attitude formation toward mobile applications. *International Journal of Electronic Commerce, 19*(4), 77-111.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically Assessing Review Helpfulness. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)* (pp. 423-430). Sydney, Australia: Association for Computational Linguistics.

Krishnamoorthy, S. (2015, May 1). Linguistic Features for Review Helpfulness Prediction. *Expert Systems with Applications, 42*(7), 3751–3759. doi:10.1016/j.eswa.2014.12.044

Kuan, K. K., Hui, K.-L., Prasarnphanich, P., & Lai, H.-Y. (2015, January). What makes a review voted? an empirical investigation of review voting in online review systems. *Journal of the Association for Information Systems, 16*(1), 48-71. Retrieved from http://aisel.aisnet.org/jais/vol16/iss1/1

Lacity, M. C., & Janson, M. A. (1994). Understanding Qualitative Data: A Framework of Text Analysis Methods. *Journal of Management Information Systems, 11*(2), 137-155.

Leigh, T. W., & Rethans, A. J. (1984). A Script-theoretic analysis of Industrial Purchasing Behavior. *Journal of Marketing, 48*, 22-32.

Li, M., Huang, L., Tan, C.-H., & Wei, K.-K. (2013, Summer). Helpfulness of online product reviews as seen by consumers: source and content features. *International Journal of Electronic Commerce, 17*(4), 101-136. doi:10.2753/JEC1086-4415170404

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., & Zhou, M. (2007). Low-quality Product Review Detection in Opinion Summarization. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 334-342). Prague: Association for Computational Linguistics.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In L. Ungar, M. Craven, D. Gunopulos, & T. Eliassi-Rad (Ed.), *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)* (pp. 935-940). New York, NY, USA: ACM.

Mudambi, S. M., & Schuff, D. (2010, March). What Makes A Helpful Online Review? A Study Of Customer Reviews On Amazon.com. (C. Saunders, Ed.) *MIS Quarterly, 34*(1), 185-200.

Ngo-Ye, T. L., & Sinha, A. P. (2012, July). Analyzing Online Review Helpfulness Using a Regressional ReliefF-Enhanced Text Mining Method. *ACM Transactions on Management Information Systems, 3*(2), 10:1-10:20. doi:10.1145/2229156.2229158

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems, 61*, 47–58. Retrieved from http://dx.doi.org/10.1016/j.dss.2014.01.011

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count, LIWC2001 Manual.* The University of Texas at Austin and The University of Auckland, New Zealand. Mahwah, NJ: Lawrence Erlbaum.

Rumelhart, D. E., & Norman, D. A. (1988). Representation in Memory. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' Handbook of Experimental Psychology* (Vol. 2, pp. 511-587). New York: Wiley.

Salehan, M., & Kim, D. J. (2016, January). Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics. *Decision Support Systems, 81*, 30-40. doi:10.1016/j.dss.2015.10.006

Schank, R. C. (1982). *Dynamic Memory A Theory of Reminding and Learning in Computers and People.* Cambridge, New York: Cambridge University Press.

Schank, R. C. (1999). *Dynamic Memory Revisited.* Cambridge, New York: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Sebastiani, F. (2002, March). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34*(1), 1-47.

Shmueli, G., & Koppius, O. R. (2011, September). Predictive analytics in information systems research. *MIS Quarterly, 35*(3), 553-572.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems, 46*, 287-299.

Wan, Y., & Nakayama, M. (2014, August). The reliability of online review helpfulness. *Journal of E-commerce Research, 15*(3), 179-189. Retrieved from http://www.jecr.org/node/439

Weiten, W. (2008). *Psychology: Themes and Variations* (8th ed.). Wadsworth Publishing, Thomson Learning, Inc.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Burlington, MA, USA: Morgan Kaufmann.

Yin, D., Mitra, S., & Zhang, H. (2016, March). When do consumers value positive vs. negative reviews? an empirical investigation of confirmation bias in online word of mouth. *Information Systems Research, 27*(1), 131-144. Retrieved from http://dx.doi.org/10.1287/isre.2015.0617

Zhang, Z. (2008, September/October). Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications. *IEEE Intelligent Systems*, 42-49.

Zhang, Z., & Varadarajan, B. (2006). Utility Scoring of Product Reviews. *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)* (pp. 51–57). Arlington, Virginia: ACM.

Zheng, X., Zhua, S., & Lin, Z. (2013, December). Capturing the essence of word-of-mouth for social commerce: assessing the quality of online e-commerce reviews by a semi-supervised approach. *Decision Support Systems, 56*(1), 211–222. doi:10.1016/j.dss.2013.06.002

Zhu, L., Yin, G., & He, W. (2014, November). Is this opinion leader's review useful? peripheral cues for online review helpfulness. *Journal of E-commerce Research, 15*(4), 267-280. Retrieved from http://www.jecr.org/node/449