International Journal of Information Management 54 (2020) 102132

Contents lists available at ScienceDirect

International Journal of Information Management

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

A comparative assessment of sentiment analysis and star ratings for consumer reviews

Sameh Al-Natour*, Ozgur Turetken

Ted Rogers School of Management, Ryerson University, Toronto, M5B 2K3, Canada

ARTICLE INFO

Keywords:
Sentiment analysis
eWOM
Consumer reviews
Machine-learning
Comparative assessment

ABSTRACT

Electronic word of mouth (eWOM) is prominent and abundant in consumer domains. Both consumers and product/service providers need help in understanding and navigating the resulting information spaces, which are vast and dynamic. The general tone or polarity of reviews, blogs or tweets provides such help. In this paper, we explore the viability of automatic sentiment analysis (SA) for assessing the polarity of a product or a service review. To do so, we examine the potential of the major approaches to sentiment analysis, along with star ratings, in capturing the true sentiment of a review. We further model contextual factors (specifically, product type and review length) as two moderators affecting SA accuracy. The results of our analysis of 900 reviews suggest that different tools representing the main approaches to SA display differing levels of accuracy, yet overall, SA is very effective in detecting the underlying tone of the analyzed content, and can be used as a complement or an alternative to star ratings. The results further reveal that contextual factors such as product type and review length, play a role in affecting the ability of a technique to reflect the true sentiment of a review.

1. Introduction

Word of mouth (WOM) refers to the personal communication between individuals concerning their perception of goods and services (Ye et al., 2009). WOM has traditionally been used by consumers to learn about products and services and evaluate their attributes prior to purchase (Chatterjee, 2019). Since the emergence of e-business, research has consistently shown that online or electronic word-of-mouth (eWOM) plays an even more important role than traditional WOM in shaping consumer attitudes and behaviors (Bulbul et al., 2014; Chen & Xie, 2008; Cheung & Lee, 2012; Dang et al., 2010; Hu & Chen, 2016; Pavlou & Dimoka, 2006).

eWOM comes in many forms such as consumer reviews, expert blogs, or microblogging sites such as Twitter. The proliferation of eWOM through these platforms generates a vast amount of content. Hence, consistent with the rules of economics of information (Nelson, 1970), finding relevant content in this collection of consumer reviews comes with a "search cost" in terms of effort and time, which hinders the utilization of eWOM. Therefore, consumers need to lower the total cost of information search by discerning what is most useful to read. Information search theories suggest that mechanisms that provide a "scent" (Pirolli, 2007) or "signal" (Payne et al., 1993) as to what a piece of information might contain can be important aids for consumers trying to decide what information to use. One aspect of such "information scent" is the "tone" or "sentiment" of a review (Mudambi & Schuff, 2010; Pavlou & Dimoka, 2006), which represents an overall evaluation of its polarity, and can be provided manually, or automatically through big data/text analytics.

In the consumer review context, star ratings are commonly used to give a quick indication of review sentiment, and are therefore the standard scent-providing decision aid accompanying reviews. However, star ratings have some limitations. In many instances, they are biased (Qiu et al., 2012). In other instances, the star rating and the review sentiment do not match (Zhang et al., 2011), such as the case when a star score is high but the review is negative or when a star score is low but the review is positive. Hence, when analyzing reviews, Valdivia et al. (2017) suggest "… not setting the user rate as a label sentiment for the whole review and analyzing the opinions in depth" (p. 75). Similarly, Kordzadeh (2019) found that there may be biases in star ratings based on where these ratings are published, lowering their reliability. Further, star ratings cannot be applied to a specific part of a document, and are typically absent in certain forms of eWOM such as blogs or tweets. These limitations highlight the potential of automatic sentiment analysis (SA), an analytics technique that assesses the tone of text, for identifying the true sentiment of a piece of text as an alternative, or a complement, to explicit expressions such as star ratings.

* Corresponding author at: 350 Victoria St., Toronto, ON M5B 2K3, Canada.
E-mail address: salnatour@ryerson.ca (S. Al-Natour).

https://doi.org/10.1016/j.ijinfomgt.2020.102132
Received 21 August 2019; Received in revised form 19 April 2020; Accepted 20 April 2020
Available online 19 May 2020

Like many text analytics tasks, SA can be performed through a knowledge heavy (i.e., lexicon based) approach that requires the creation of extensive repositories, or a knowledge-light (machine learning) approach (Feldman, 2013; Roussinov & Turetken, 2009) that requires training moels with large data sets with minimal manual intervention. There are numerous tools readily available that implement many variations of these general approaches. Nevertheless, there is no adequate insight as to how the performance of the main approaches to SA compare to each other and the almost omnipresent star ratings for a particular type of review. This is the main gap in the literature that we aim to fill. As such, the focus of the paper is not to add to the extensive body of work that focuses on improving SA algorithms (see Feldman, 2013, for an insightful review). Rather, we aim to comparatively assess the ability of the state of the art tools that represent the main approaches to sentiment analysis in identifying the true sentiment of a consumer review.

Past comparisons of different approaches to SA (Annett & Kondrak, 2008; Taboada et al., 2011) produced findings that are often ungeneralizable. One reason for this is that researchers typically compared tools that implement variations of the same technique, for example Support Vector Machine (SVM) or Naïve Bayes (in machine learning tools), or those that calculate sentiment scores through similar approaches but using different lexicons (in lexicon-based tools). A methodic comparison of fundamentally different techniques, for example comparison of lexicon-based approaches to machine learning approaches, is extremely rare (see Kirilenko et al., 2018 as an exception). Moreover, the performance of tools based on more recent techniques such as deep learning, which represents a strong trend in modern machine learning, is not known (Valdivia et al., 2017). In contrast, in this study, we conduct a systematic comparison between the most common approaches (or techniques) to SA. In so doing, we attempt to evaluate the performance of these approaches as an alternative, and a complement, to star ratings. The pervasiveness of star ratings makes it worthwhile to study their utility in an environment where multiple alternatives to SA are available.

Another limitation in past research is the criteria used to compare different approaches. In this paper, we model sentiment scores as a continuous variable. As such, we are able to assess SA accuracy in a finer fashion than what has been done in the past, where many studies treated SA as a two-way or three-way classification problem (e.g. Hasan et al., 2018; Kirilenko et al., 2018; Zhang et al., 2011) that categorizes text into the broad categories of negative, positive or neutral. Such categorization results in losing a certain amount of variance that is only possible to account for when sentiment is modeled as a continuous variable. Modeling SA scores as a continuous variable also makes it possible to combine the output of multiple tools (including star ratings), whereas reconciling the scores from SA tools that are merely (negative, neutral, positive) classifiers is a much bigger challenge (Kirilenko et al., 2018). This, as we discuss later, is an advantage to the approach we took, and a contribution of our work.

We also consider contextual factors that have been largely ignored in the SA literature. The first contextual factor is review length, which is worthy of examination considering that the majority of SA studies focus on analyzing short text, often in the form of tweets (e.g., Saif et al., 2016), which limits their generalizability. In contrast, consumer reviews vary greatly in length. Studies examining the accuracy of various SA tools, have shown that the accuracy of tools can differ greatly when analyzing individual phrases, sentences, or complete documents (Agarwal et al., 2011; Wilson et al., 2005). This suggests that review length should moderate how effective SA tools are in detecting sentiment. Yet to our knowledge, prior research has not compared SA scores for textual content with varying length.

Product or service type has been identified as another contextual factor in consumer decision-making (De Maeyer, 2012; Mudambi & Schuff, 2010; Philander & Zhong, 2016; Zhu & Zhang, 2010). Consumers' approach to decision making differs depending on the type of product or service under consideration. This is also reflected in the characteristics of eWOM that is generated about these products and services. Consequently, product context can affect the accuracy of different SA approaches. As suggested by Liu (2012): "…words and even language constructs used in different domains for expressing opinions can be quite different." For example, the same word may be perceived to be positive in one domain, but negative in another (p. 38). In consumer reviews, the product or service type will likely affect the objectivity/subjectivity of the text and the choice of language constructs. Given the differences in how SA techniques handle these variations, it is likely that product or service type moderates the accuracy of an SA technique. Yet, a systematic comparison of SA tools in different product/service domains has not been conducted.

Understanding the factors that affect SA accuracy, be them contextual or method-based, is essential to its continued and future development and utilization. In essence, understanding these factors has the potential not only to inform how the SA algorithms are designed (algorithm design being the most researched area in SA), but also how and where to effectively apply SA tools (i.e., their utility). Consequently, this study focuses on investigating the potential of SA in general, and the factors that affect its accuracy in particular, and attempts to answer the following questions:

1 How effectively can SA scores generated by different SA approaches (as complements or as alternatives to star ratings) predict the true sentiment expressed in a product/service review?
2 How do SA scores generated by different approaches (including star ratings) compare to, and complement, each other in the context of different product types and review length?

The remainder of this paper is organized as follows. The following section provides a theoretical background, offers a brief review of the literature on sentiment analysis and its various tools, introduces the context within which these tools are used and compared, and identifies gaps in the literature. Section 3 presents our research model and hypotheses development. The two subsequent sections describe the methodology used in the study, and present the study's results. This is followed by a discussion of these results, the study's limitations, directions for future research, and concluding remarks.

2. Background

2.1. Sentiment analysis and information search

Sentiment analysis has been used for a range of purposes such as tracking the popularity and desirability of a brand (e.g., Greco & Polli, 2019), identifying product opportunities (e.g., Jeong et al., 2017), studying product launches (e.g., Rathore & Ilavarasan, 2020), predicting market movement of stocks (e.g., Maqsood et al., 2020), detecting public sentiment for policy and political positions (e.g., Wu et al., 2019), planning for disaster response and recovery (e.g., Ragini et al., 2018), and as a decision aid for purchasing products and services (e.g. Feldman, 2013). As our interest is in the role of SA as a decision aid that helps consumers focus on potentially relevant information, the theoretical bases for framing this research comes from information search, namely information foraging theory (Pirolli, 2007).

According to information foraging theory (Pirolli, 2007), the mechanisms within which people search for information resembles the way animals search or forage for food in the wild. The two main tenets of foraging are the aid the foragers use in locating food (or information), and how they form their diet by moving from one food (or information) patch to another. For our purposes, the former of these is relevant. Information foraging theory posits that akin to the use of smell by animals foraging food, individuals use "information scent" as a signal as to where there might be relevant information that they can subsequently consume. Therefore, when large amounts of consumer

reviews are available, it is important that information scents are provided since searching for the most relevant information without adequate guidance can be prohibitive.

Accordingly, research has shown that providing cues regarding review polarity is seen to be useful by consumers (Mudambi & Schuff, 2010), i.e. review polarity is a useful information scent. On online shopping portals, this is typically accomplished by using star ratings, which, as discussed earlier, have a number of limitations. This raises the question of whether SA scores can substitute for star ratings. The answer to this question is to a large extent dependent on whether SA can accurately reflect the true sentiment of the analyzed text, as both the actual and perceived value of SA tools will depend on their accuracy. Hence, validating SA accuracy becomes a prerequisite for the utility of SA for eWOM. This paper focuses on this point by systematically assessing the accuracy of different SA approaches, along with star ratings.

2.2. Theoretical foundations of sentiment analysis

Sentiment analysis utilizes automatic natural language processing techniques to detect the sentiment of a given piece of text to calculate an overall sentiment score (Liu, 2010, 2012; Nasukawa & Yi, 2003; Pozzi et al., 2016; Taboada et al., 2011). The two major approaches to sentiment analysis are lexicon-based (linguistic) and machine learning (Pradhan et al., 2016). Both approaches have the potential to produce a categorical (i.e. positive/neutral/negative) or continuous sentiment score. Lexicon or dictionary-based techniques rely on word-maps such as SentiWordNet (e.g. Ribeiro et al., 2016), which contain four sets of terms: positive, neutral, negative, and stop terms. The strength of the positive and negative terms is predetermined, e.g. “great” is more positive than “good”. After stop terms are eliminated, the remaining text is checked for the frequency of positive, negative, and neutral terms using lookup algorithms (Chopra et al., 2016). Lexicon-based tools (e.g. Hutto & Gilbert, 2014) typically revise the original sentiment of these terms based on modifiers such as “very”. The overall score for the text is then calculated based on these revised polarity scores. When the output is categorical, the score is determined by whichever sentiment has the highest frequency. The overall sentiment can also be represented on a continuous scale as the net polarity of the text after the polarity of the terms are added together (i.e. sum of the scores of the positive sentiment terms minus those of the negative terms). Some tools adopt a variance of this approach, such as calculating the ratio of positive, negative or neutral terms to the total number of sentiment terms (Philander & Zhong, 2016).

The machine learning (ML) approach treats sentiment analysis as a pattern recognition problem using established techniques for classification or prediction (Abe, 2005). The main advantage of ML-based SA is that it does not rely on word dictionaries, which may be very costly to create and maintain (Dang et al., 2009; Ye et al., 2009). Instead, a broader set of features are extracted from data, which are more comprehensive than typical sentiment terms (i.e. adjectives), and include nouns representing objects and verbs representing attitudes towards those objects (Liu, 2012). However, ML based techniques need labeled data sets for supervised training, which has been identified as their shortcoming (Ribeiro et al., 2016).

Since most research treated SA as a classification problem, the majority of ML-based techniques reported in the literature are classifiers (using well-known methods such as support vector machines (SVM) or Naïve Bayes classification, the algorithmic details of which are beyond the scope of this research) that use data labeled with a group membership among positive and negative (-1, 1), or positive, negative, or neutral (-1, 0, 1) categories. There have been other techniques used for the SA problem, most notably, neural network (ANN) based models. ANN models mimic the way a human brain works, where artificial neurons process input through a nonlinear transformation and pass their output to neurons that they are connected to in the next layer of the network (Nielsen, 2015). The neurons in the input layer of an ANN-based SA tool receive the document features as their input, whereas the input to the neurons in the further (deeper) layers of the network is the sum of the outputs of the neurons from the previous layer, which are weighted by the strength of the inter-neuron links. An ANN can have many layers (depth) depending on the complexity of the relationship between the inputs and the output. The output of the neuron or neurons at the deepest (output) layer of the ANN-based SA tool determine the sentiment score. ANN-based models generate a continuous output which can then be categorized for categorical sentiment scores. The ANN output (i.e. predicted score) is compared to the known sentiment of the input document and the difference is propagated backwards in the network, which is then used to modify the strength of the links between neurons in consecutive layers of each other. This process, known as the training of the network, continues until the difference between the predicted and the actual output (the error) has stabilized. Network training can be a lengthy process especially for deep networks that have more than one layer between the input and output layers. As such, the high computational cost of training ANN was reported as a disadvantage although these models can result in superior performance over those built based on techniques such as SVM (Moraes et al., 2013).

Theoretically, the lexicon-based approach to SA is expected to exhibit an inherent advantage in simulating the effects of linguistic context since aspects of the local context of a word are taken into account when building a lexicon for a specific domain (Taboada et al., 2011). The main disadvantage of lexicon based techniques on the other hand is the high computational cost of building a lexicon. Similarly, ML-based techniques can also be considered domain specific if the training data set represents a specific domain. If this reduces the ability of a model to predict sentiments in an entirely new domain, then the model would have to be re-trained for different domains with new data sets at a high computational cost. Another disadvantage of ML based approaches is that the training domain is embedded in the models. The algorithm therefore is opaque and virtually impossible to understand by its users, whereas a lexicon can be more easily studied and modified if necessary. On the other hand, ML-based techniques have the advantage of being flexible and able to improve their prediction models as new data and findings become available.

To summarize, there are numerous reasons why a specific approach to SA, and hence its representative applications (i.e. tools), would be effective in comparison to the alternatives. It is virtually impossible to make these assessments solely based on theory, which makes empirical comparisons essential. Next is a review of that stream of research.

2.3. SA comparison studies

Past literature, especially in the tourism domain where SA has gained a lot popularity (Schuckert et al., 2015), reports on many studies where SA approaches (or tools that represent these approaches) are compared. Ye et al. (2009) compared three supervised machine learning algorithms of Naïve Bayes, SVM, and the character based N-gram model for sentiment classification of reviews on travel blogs for seven popular travel destinations in the US and Europe. Their findings suggest that the SVM and N-gram approaches are superior to the Naïve Bayes approach. However, when the training datasets had a large number of reviews, all three techniques reached more than acceptable levels of accuracy (at least 80 %). This is one of the earlier studies in the literature that points to model training challenges as a general disadvantage of ML approaches.

Moraes et al. (2013) compared SVM and ANN for document-level SA where sentiment scores were modeled as a binary variable. The findings indicated that ANNs produce superior or at least comparable results to SVM, in some cases, by a statistically significant difference. The authors emphasize that the limitation of SVM is the computational cost at the running time, while the limitation of ANN is the computational cost at the training time. Due to this limitation, ANN had been rarely used in

SA comparisons up to that point in time, which is still true today.

Hutto & Gilbert (2014) introduced VADER (Valence Aware Dictionary for sEntiment Reasoning), which utilizes human expertise in building a rich lexicon. It performed comparably to, and in many cases better than, the leading lexicon-based SA tool of the time LIWC, along with six other well-established tools. As noted by its creators, VADER is specifically attuned to sentiment in microblog-like contexts (Hutto & Gilbert, 2014). Nevertheless, it has been considered the gold-standard of lexicon-based SA for numerous types of textual content (e.g. Ribeiro et al., 2016).

Gao et al. (2015) examined three web services that provide SA functionalities. The comparison of AlchemyAPI, Semantria, and Text2Data on hotel data revealed that all of the tools exhibit high levels of accuracy for positive reviews, but not for neutral ones. The authors concluded that their results provide users with guidance on what SA tool to use for what context. However, their limited elaboration on the criteria for the selection of these tools and on the contextual parameters of the dataset limits the contribution of this work. Likewise, Valdivia et al. (2017) compared SentiStrength, Bing, Syuzhet, and CoreNLP6 on TripAdvisor data, and found that SentiStrength and Syuzhet more closely approximated the distribution of star ratings while Bing and CoreNLP did not. In a similar study, Hasan et al. (2018) compared three lexicon-based tools, namely SentiWordNet, TextBlob, and W-WSD on twitter based data to detect sentiments concerning political parties and candidates, and found that overall TextBlob leads to the best classification accuracy. Similar to the case in Gao et al. (2015) and Valdivia et al. (2017), the seemingly random selection of the SA tools is the limitation of this work.

Kirilenko et al. (2018) compared a lexicon-based tool SentiStrength with the ANN based Deeply Moving, and two other ML based tools (SVM and Naive Bayes) developed by the authors in the RapidMiner environment. The tools were tested on three sets of tourism related data. The results showed that both SVM and Naive Bayes outperformed the other two off-the-shelf tools. Nevertheless, the authors conclude that an off-the-shelf lexicon-based approach is preferable to an off-the-shelf machine learning algorithm if no training or calibrating is feasible due to time or cost constraints.

There have been attempts to use SA approaches as complements to each other. Prabowo & Thelwall (2009) combined rule-based classification, supervised learning and unsupervised machine learning into a new combined method, and tested this method on movie reviews, product reviews and MySpace comments. The results show that hybrid classification can improve the classification effectiveness in terms of F1, which takes both the precision and recall of a classifier's effectiveness into account. In addition, they tested a semi-automatic, complementary approach in which each classifier can contribute to other classifiers to achieve a good level of effectiveness. Similarly, Appel et al. (2016) introduced a hybrid method, which led to results that are more accurate and precise than both Naïve Bayes and Maximum Entropy when these techniques are utilized in isolation. Because these studies modeled SA as a classification problem, the combination of the techniques had to be performed at the algorithm level. In fact, some researchers specifically advised against combining classification scores (Kirilenko et al., 2018). On the contrary, continuous SA scores allow for the simple combination of scores without creating new hybrid algorithms, which is the approach that we followed in this research. To our knowledge, such an approach is novel.

2.4. Research gaps

Our study differs from past empirical SA studies, and contributes to the literature in a number of ways. First, a vast majority of the approaches reported in the literature treat SA score as a categorical variable taking on a value of positive, negative, or neutral (e.g. Gao et al., 2015; Hasan et al., 2018; Kirilenko et al., 2018; Liu, 2012; Zhang et al., 2011). Similarly, the majority of comparative studies (e.g. Annett & Kondrak, 2008; Ribeiro et al., 2016; Taboada et al., 2011; Ye et al., 2009) compare different SA alternatives on sentiment classification, or metrics derived from these such as precision, recall, or F scores. However, there are advantages to modeling sentiment scores on a scale ranging from extreme negativity to extreme positivity (Liu, 2012). For example, as discussed by Philander & Zhong (2016), SA can be used to capture changing sentiments on a topic of interest; a task that is much easier to do if the slight changes in sentiments are captured before they are large enough to move from one category (positive/negative/neutral) to another. If SA scores are to be modeled as a continuous variable, negative trends in opinions regarding a product can be captured before the sentiment moves from positive to neutral or neutral to negative.

Past research has also used previously labeled datasets (e.g. Annett & Kondrak, 2008; Taboada et al., 2011) or star ratings (Ye et al., 2009) as bemchmarks for accuracy. Ribeiro et al. (2016) used human scorers (generated from MechanicalTurk) as we do in this study, yet, to our knowledge, ours is the only study so far that models sentiment scores as a continuous variable while generating original scores using independent human judges. Therefore, we have a high level of granularity in our sentiment assessments. Continuous sentiment scores also make it possible to combine outputs from different tools more conveniently without the need for further algorithmic development. This is in line with the suggestions of Valdivia et al. (2017), which advocate trying to obtain consensus among SA scores.

To the best of our knowledge, ours is also the first attempt to comparatively evaluate SA tools (both lexicon and ML-based) along with star ratings. In fact, many studies examining the utility of SA (Ribeiro et al., 2016; Taboada et al., 2011), have explicitly excluded not only star ratings, but focused only on lexicon or ML-based tools at a time. We also are interested in tools that every day decision makers can readily use; therefore, our comparison focuses easy to use software packages that do not require additional coding. In that sense, our approach resembles that of Ribeiro et al. (2016), and departs from the majority of comparison studies that report on models developed by the authors themselves (e.g. Annett & Kondrak, 2008; Taboada et al., 2011; Ye et al., 2009), which are not readily available to everyday users. In fact, Kirilenko et al. (2018) advise against the use of off-the-shelf ML tools due to the aforementioned model training challenges, especially for ANN, which perform better otherwise (Moraes et al., 2013). As described later in the paper, for our study, we use one of the most recent ANN-based SA tools, which is pre-trained with very large datasets, along with gold standard lexicon based (pure and hybrid) tools. As such, our results provide a good glimpse of the state of the art in this field.

Previous assessments of SA approaches also lack a systematic inclusion of contextual factors. Review length and product/service type have been identified as determinants of review helpfulness (e.g., De Maeyer, 2012; Mudambi & Schuff, 2010; Zhu & Zhang, 2010). We believe those variables are relevant for SA research as well, since it has been shown that SA accuracy can differ greatly when analyzing individual phrases, sentences, or complete documents (e.g., Agarwal et al., 2011), which suggests that length of the analyzed text is a factor that affects the accuracy of SA scores. Similarly, domain can change the meaning of words used, and hence the ability of SA tools to accurately capture the sentiment expressed through them (Liu, 2012). Since domain in consumer reviews is largely defined by the type of good addressed in the review, product/service type is another factor that influences SA accuracy. Past research has not made a conscious effort to account for these contextual variables. As a consequence, the review data the tools were applied to were typically unbalanced and not generalizable across varying review domains. To address this particular gap, we employ a factorial research design that accounts for consumer reviews of varying length that concern various product contexts, while ensuring that there are multiple products for each type/length combination. This sets this research apart from previous comparison studies such as Annett & Kondrak (2008), which only considered a movie

dataset, or Ye et al. (2009), which only considered data on travel destinations. Taboada et al. (2011) report on similarly diverse data sets yet their sample is not balanced in terms of product category or length. Likewise, Ribeiro et al. (2016) use a variety of databases including social network data and tweets, yet without a theoretical classification of the review content.

3. Research model and hypotheses

Past literature has established that there is merit in capturing review sentiment through star ratings or with SA tools. What has not been fully established though is whether these ratings or SA scores have marginal value, i.e., whether scores obtained from different SA tools and star ratings can add a unique contribution to our ability to identify the true sentiment of a piece of text (e.g., customer review). The distinct advantages and disadvantages of different SA techniques and star ratings that we have reviewed lead us to conclude that there is reason to answer this question in the affirmative. Given the fundamental differences between these techniques in how ratings and sentiment scores are generated and quantified, they should share minimal variance. In other words, the differently-generated sentiment scores as well as star ratings should be largely independent of each other, and therefore their errors when estimating the true sentiment of a review (score error) should not be correlated. Hence, we propose that each of these scores will contribute uniquely to the prediction of the true sentiment of a review.

H1. Sentiment scores that are based on different approaches act as significant predictors of the true sentiment of a review.

While prior research has recognized the potential of different SA tools to act as complements of each other (Appel et al., 2016; Prabowo & Thelwall, 2009), no studies have looked at the potential enhancements that can be obtained by simply combining the outputs of existing tools (especially those that are available off-the-shelf). Instead, the focus of past research has been on modifying the algorithms that form the basis of these tools, in order to create hybrid tools that embody selective choices (Alaei et al., 2019). In this study, we propose that combining the outputs of individual tools can be a more fruitful approach than attempting to amalgamate their features. Hence, and as a corollary to hypothesis H1, we further propose that a combination of individual SA scores is more predictive than individual ones.

Prior research has recognized the potential of SA tools that are based on different approaches to generate uncorrelated and largely divergent scores (e.g., Liu, 2012). As described earlier, given the variations in how the main approaches quantify the sentiment expressed in textual content, the generated scores (including star ratings) should be largely independent. Hence, the errors in these scores when estimating the true sentiment of a review (score error) should not be correlated, and each score should predict unique variance in the true sentiment of a review. Consequently, a score that combines the individual scores should account for those unique contributions and improve the sentiment prediction accuracy over each individual score. More formally:

H2. The accuracy of combined sentiment scores when predicting the true sentiment of a review will be higher than the accuracy of individual scores.

Past research found that review depth (length) improves diagnosticity and affects perceived helpfulness (Ghasemaghaei et al., 2018; Mudambi & Schuff, 2010). Short reviews often lack a comprehensive assessment of product features, while longer reviews often contain deeper analyses of the product and more emotions (Ghasemaghaei et al., 2018). As such, shorter reviews would likely be more direct, i.e. only emphasize the overall sentiment of the review, as the reviewer is unlikely to offer detailed assessments of the various dimensions of the product/service being reviewed. Meanwhile, longer reviews are likely to be more nuanced and detailed, which may lead the review to include both positive and negative sentiments with different levels of emphasis on each.

Notwithstanding this effect of review length on helpfulness and comprehensiveness, we propose that length would be a moderator of the relationship between an SA approach and the accuracy of sentiment scores. For lexicon-based tools, when scores for sentiment words are averaged to obtain an overall score for the whole review, the abundance of both positive and negative sentiment terms would lead the overall score to exhibit a tendency towards more neutral scores for long reviews, and as a result lowering accuracy. There is support for this argument in the literature (e.g., Kirilenko et al., 2018), where lexicon-based tools were shown to perform better for shorter text. However, this is an *a posteriori* finding as review length was not explicitly manipulated in that study. Likewise, partial evidence comes from other research that also did not explicitly account for review length (Lak and Turetken 2014), which found that lexicon-based SA scores deviated most from star ratings for extreme reviews. Last, but not least, some of the most popular and successful lexicon-based tools are optimized for sentiment detection in microblog-like contexts (e.g., Hutto & Gilbert, 2014). Hence, we expect that their accuracy will also be optimized for similar short text. Therefore:

H3a. Review length moderates the accuracy of lexicon-based SA tools such that the accuracy is higher for short reviews then medium reviews then long reviews.

Despite their potential for improving classification accuracy (Appel et al., 2016; Prabowo & Thelwall, 2009) over a pure lexicn-based tool, hybrid tools that employ a lexicon-based approach, even in part, are likely to be similarly deficient when analyzing longer text. Regardless of the specific type of approach with which the lexicon approach is combined, the tool will still use the sentiment of individual phrases to obtain an overall score for the whole review. A longer review would likely include numerous sentiment phrases with varying polarity. Without a contextual synthesis of those phrases, a hybrid tool would suffer from the shortcomings of pure lexicon-based tools when it comes to longer reviews. Hence, we formulate the following hypothesis:

H3b. Review length moderates the accuracy of hybrid (ML/lexicon) SA tools such that the accuracy is higher for short reviews then medium reviews then long reviews.

In contrast, techniques based on human judgement, i.e. star ratings and machine learning techniques, specifically the ANN-based deep learning algorithms that we examine in this research, are known for their pattern recognition capabilities, which should not be impacted by document length and nuances as in the case of lexicon-based techniques. Therefore, we expect the performance of these scores to be more robust and not moderated by the length of the text analyzed.

Another important aspect of the context in purchase decisions is the category of the product or service that is being reviewed. Consumer research has identified the following three categories of products/services: 1) *search goods*: which can be evaluated prior to purchase (e.g., electronics); 2) *experience goods*: which can be evaluated only after purchase (e.g., a hotel room); and 3) *credence goods*: which the consumer can never completely evaluate even after the purchase (e.g., multivitamins) (Mills & Law, 2004). By their very nature, search goods are amenable to more objective evaluations and fewer sentiments (Feldman, 2013). As such, it is more likely for search product reviews to emphasize the overall sentiment of the review with less nuance.

Given that SA tools analyze what is identified as the subjective part of text, the performance of SA tools in analyzing search products should resemble that for short reviews. Consequently, we expect lexicon based tools to produce more accurate sentiment scores for search goods as they do for short reviews. On the other hand, reviews concerning experience and credence goods are likely to contain subjective evaluations, and hence reviews for experience and credence products should resemble those of medium-length and longer reviews. Specifically, for lexicon-based tools, when scores for sentiment words are averaged to

obtain an overall score for the whole review, there would be a tendency towards generating more neutral scores for more subjective reviews, which would lower accuracy. Therefore:

H4a. Product type moderates the accuracy of lexicon-based SA tools such that the accuracy is the highest for search products then experience products then credence products.

As described earlier, hybrid tools that apply a lexicon in their processing, will suffer from the same shortcomings of pure lexicon-based tools. More specifically, reviews for experience and credence goods are expected to be more subjective and nuanced. As such, they would likely include numerous sentiment phrases with varying polarity. Without a contextual synthesis of those phrases, the performance of a hybrid tool is likely to be impacted by the subjectivity of the content in experience and credence product reviews. Hence, we expect a moderation effect of product type, similar to what would be observed for pure lexicon-based tools, would also apply to hybrid tools.

H4b. Product type moderates the accuracy of hybrid (ML/lexicon) SA tools such that the accuracy is the highest for search products then experience products then credence products.

As in the case of review length, product type should not significantly impact the performance of ML based tools or star ratings. In fact, Kirilenko et al. (2018) found that humans' assessment of textual documents showed little variance across different datasets, which is seen as an advantage of decision aids such as star ratings. This suggests a domain independent nature of star ratings that are a direct product of human intelligence, and to a degree, machine learning tools that involve ANN/deep learning that mimic human intelligence.

Fig. 1 depicts our research model that summarizes the above discussion and resulting hypotheses.

4. Method

4.1. Sample

To test the hypotheses, we first collected a dataset of 60,728 consumer reviews from Amazon and Yelp (30,337 reviews from Amazon and 30,182 from Yelp). The reviews concerned six products/services that can be categorized into the three groups of search, experience, and credence goods. For search products/services (20,272 reviews), we selected reviews of laptop computers and digital cameras. All search product reviews were extracted from Amazon. For experience products/services (20,433 reviews), the reviews concerned hotels and restaurants. All search product reviews were extracted from Yelp. The two selected credence products/services were auto repair (9749 from Yelp) and multi-vitamins (10,065 from Amazon). The selected products/services ranged in value and price, reflecting different levels of involvement on the part of the reviewers. All the reviews used a rating scale from 1–5 stars, and ranged in length from 5 to 1717 words.

The Amazon reviews were extracted from Amazon.com using the Amazon Web Services Product Advertising API. To select the reviews, we extracted the first 10,000+ reviews that address any of the products of concern (e.g., digital cameras). The Yelp reviews were similarly extracted using the Yelp Fusion API, which gives developers access to Yelp's local content concerning 50 million businesses across their 32 international markets. To select the reviews, we extracted the first 10,000+ reviews that concern hotels, restaurants or auto repair services located in Toronto, Canada. For each review, we collected the review text, the product/service type, and the website on which the review is displayed.

To categorize the reviews in terms of length, we performed a frequency analysis. The results indicated that approximately one third of the reviews in the dataset are shorter than 40 words, while another third is 100 words or more. Consequently, we categorized *short* reviews as those with a word count less than 40 words, *medium* reviews with a count of 40–99, and *long* reviews with 100 words or more.

Since our objective is to comparatively evaluate star ratings and SA scores in identifying the true sentiment of reviews, we selected reviews balanced for product/service type and length to be evaluated by human judges. To select the reviews, we first categorized the larger set of 60,000+ reviews based on the product the review addresses (6 levels), the length category of the review (3 levels), and the star rating given by the review's author (5 levels). This categorization resulted in 90 different sub-classifications (e.g., short digital camera reviews with 1 star; short digital camera review with 2 stars' … etc.). We then randomly selected 10 reviews from each subset. The cell size of 10 gave us a sufficient total sample size of 900 (90 × 10) for statistical power considerations, while also keeping the manual assessment of these reviews feasible. The balanced selection ensured that the selected reviews are representative of the whole spectrum of opinions, as the eventual sample is balanced in terms of star ratings (i.e., the sample has an equal

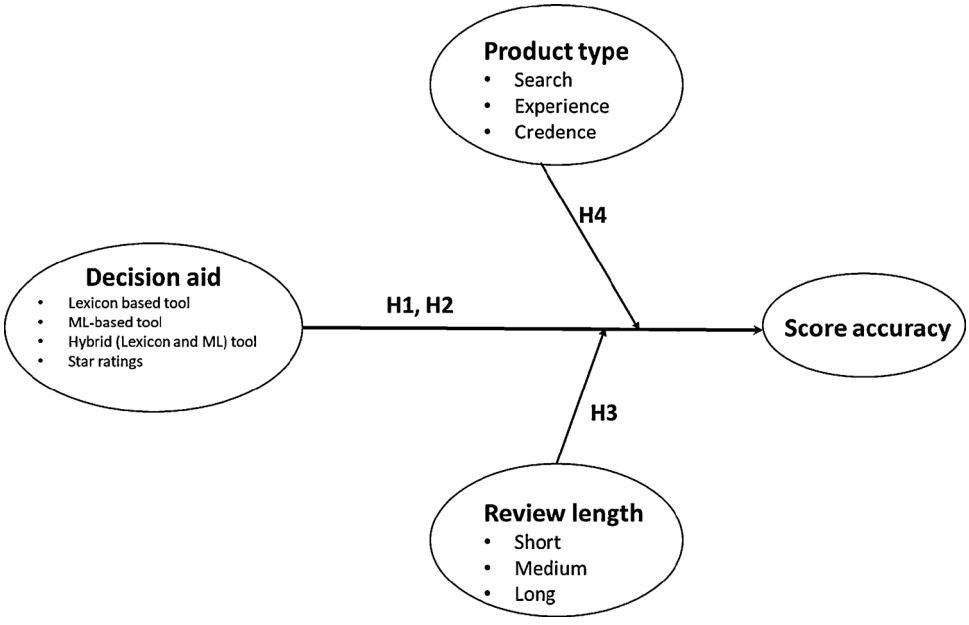

Fig. 1. The Research Model.

number of 1-star, 2-stars … etc. for each *product/service type* of each *length category*).

4.2. Generating the sentiment scores

Since the aim of the study is not to design or develop a new sentiment analysis tool but rather evaluate the state of the art off-the-self tools, we used such tools that employ lexicon-based, ML-based, and hybrid techniques. This approach is consistent with prior studies that employed an off the shelf commercial tool to represent a given approach (e.g. Schumaker et al., 2012).

To select these tools, we surveyed a large number of available tools as well as research that employed such tools. Given that the length of consumer reviews is not constrained to a specific number of characters/words, it was essential that the tools chosen have the ability to analyze text that could potentially exceed a thousand words. This criterion reduced the number of potential tools significantly. Second, given that our objective is to generate an overall score for the whole review, and not scores for specific words used or sentence-level scores, it was important that the chosen tools have the ability to perform document-level analysis (Feldman, 2013; Pradhan et al., 2016), i.e. receive a complete review as input and produce one score to represent the polarity of the whole review as an output. Third, since consumer reviews could potentially include many slang words, typos, punctuation intensifiers, emoticons, or any other types of heuristics that can be used to infer the polarity and intensity of a written text, we considered tools that consider such heuristics as part of their processing to determine polarity. Lastly, as mentioned before, to be able to compare and/or combine different scores generated by the different tools and star ratings, we chose tools' whose outputs are numeric and continuous in nature.

Applying the above criteria led to the selection of three tools. The first is VADER (Valence Aware Dictionary for sEntiment Reasoning) (Hutto & Gilbert, 2014), which is a rule-based sentiment analysis tool that has the ability to analyze text of any length to generate a numeric valance (polarity) score. The authors built a lexicon that was mostly based on pre-existing lexicons, but extended with emoticons and acronyms typically used in social media. This lexicon was subsequently reduced to a rather parsimonious set of roughly 7500 lexical features through validation by crowd sourced human coders. In addition to the polarity scores assigned for each feature in its lexicon, the VADER algorithm also incorporates rules that modify the original polarity of the lexical features based on the context in which they are used in a sentence, e.g., with (or without) punctuation marks, or in all capital letters. VADER uses these modified scores for all the lexical features present in a piece of text and averages these scores to arrive at an overall sentiment score for that text. Although designed with a special emphasis on social media, the VADER algorithm has the ability to analyze text of any length.

VADER has been validated using movie and product reviews as well as newspaper readers' opinions, and has been used as a benchmark lexicon tool in a number of studies (Alaei et al., 2019; Hutto & Gilbert, 2014; Ribeiro et al., 2016), where it was shown to be the most consistently well performing lexicon-based tool (compared to over twenty tools across numerous databases). Hence, VADER can be considered the gold standard for lexicon-based SA.

The second tool we used is Google's Cloud Natural Language API, henceforth referred to as Google (https://cloud.google.com/natural-language/). It uses machine learning, more specifically Google's deep learning platform, which employs sophisticated multi-layer (i.e. deep) ANN models to reveal the structure and meaning of text. The main disadvantage of ANN-based SA, as we mentioned in our review of the literature, is the time required for the training and re-training process. This would be especially true for deep learning networks, yet Google Natural Language API eliminates this disadvantage as it provides powerful "pre-trained" models including those for sentiment analysis, which work rather fast on a typical PC. Unlike VADER, Google's pattern (in our case sentiment) recognition algorithm is opaque (black box) to users, as it is the case with virtually all machine learning techniques both in SA and beyond. While this would be a disadvantage for those attempting algorithmic improvements to state of the art tools, it is not so for the purposes of our research where the emphasis is on the utility of available tools, and not on algorithm development.

The Google tool provides entity-level and document-level sentiment analysis; the latter of which is relevant to our research. This tool is free for data sets up to 5000 data points, so interested researchers can duplicate our results. Similar to VADER, Google outputs a numeric polarity, and has the ability to analyze complete reviews of various lengths. To the best of our knowledge, ours is the first study that specifically examines the accuracy of the Google tool, or more broadly, a deep learning based approach to SA.

The third and last tool we used is TextBlob (https://textblob.readthedocs.io/). It is a Python library, and includes an API for common natural language processing tasks, such as tagging, noun phrase extraction, sentiment analysis, classification, and translation. Its sentiment analysis feature evaluates the polarity of textual content of any size, and returns a numeric polarity score (Giatsoglou et al., 2017). Similar to VADER, TextBlob applies heuristics to analyze textual properties and characteristics to modify the original valence of the terms in its library based on textual context. Its main difference from VADER is that it uses a machine learning approach to calculate document level sentiment scores from these modified scores. Hence, it is a hybrid tool that applies both machine-learning and lexicon-based techniques. Similar to VADER, TextBlob has been extensively validated by previous research (Thavasimani & Missier, 2016; Vijayarani & Janani, 2016; Wang et al., 2015).

As described above, both Google and TextBlob offer easy-to-use APIs. Hence, to generate the SA scores using these two tools, we first developed a Python-based environment that allows us to make the appropriate calls through the API while providing the textual reviews for analysis. The APIs returned the SA scores generated by Google and TextBlob, which ranged from -1 (extremely negative) to +1 (extremely positive). To analyze the reviews using VADER, we installed the open-source sentiment tool, and made the appropriate calls through Python while providing the textual reviews for analysis. Similar to the other two tools, VADER returned the SA scores, which ranged from −1 to +1, in a text file.

4.3. Determining the true sentiment of a review

Three human judges evaluated each of the 900 randomly selected reviews. Given that native speakers of English have been shown to be better at detecting the true sentiment in a review (Pang & Lee, 2005), the judges selected were native speakers. To avoid evaluator fatigue, the judges evaluated the set of 900 reviews in four separate batches of 225 (the 900 reviews were randomly distributed amongst the four batches). Each set was provided to the judges separately, and no time limits were imposed for their return. On average, it took each judge approximately 3 days to rate each set of reviews. Each judge was asked to rate every review from 0 (extremely negative) to 100 (extremely positive). They were only provided with the textual review as well as the product/service the review concerns, but not the URL of the review or the star rating assigned by the original reviewer.

To assess the agreement between the three judges, we computed the intraclass correlation (ICC) between their scores (Griffin & Gonzalez, 1995). ICC measures absolute agreement between separate evaluations, and is used to assess the consistency, or conformity, of measurements made by multiple observers measuring the same quantity (Koo & Li, 2016). While Cohen's Kappa is often reported as the measure of inter-rater reliability in similar comparative SA studies, this measure can only be used when raters are asked to classify reviews (or whatever items are being categorized) into binary or categorical measures (e.g.,

positive and negative). However, in our case, judges are asked to rate reviews on a scale from 0 to 100 (essentially a continuous scale). Hence, the Kappa metric is not appropriate. Instead, we used ICC as it is specifically designed to calculate the absolute agreement between raters, and especially when there are more than two raters (as in this case). Unlike a Pearson correlation where a perfect linear correlation can be achieved if one rater's scores differ (by a nearly consistent amount) from another (even though not one single absolute agreement exists), an ICC gives a composite of intra-observer and inter-observer variability. Hence, a perfect ICC means that the raters' scores are exactly the same. The ICC between the scores from the three judges was 0.93, which indicates an excellent level of agreement (Cicchetti, 1994).[1]

5. Results

5.1. Predictive power of SA scores

Hypothesis 1 proposes that individual SA scores and star ratings are significant predictors of the true sentiment expressed in a review. To test this hypothesis, we performed two regressions. The first included the SA scores generated by the three examined tools as predictors of the average score assigned by the human judges, and the second included the same scores generated by the three tools along with the star ratings assigned by the reviewers as an additional predictor. Given that the star ratings for the reviews ranged from 1–5 while the sentiment scores generated by the three tools ranged from −1 to +1, both types of scores were transformed to a scale ranging from 0% (equivalent to −1 SA score or 1 star) to 100 % (equivalent to +1 SA score or 5 stars). The results of the two regressions are shown in Table 1. The results of regression #1 (which does not include star ratings) indicate that the three SA tools can collectively predict 57.3 % of the variance in the human judge scores, with Google making the largest contribution ($\beta = 0.401$, $p < 0.01$), followed by VADER ($\beta = 0.330$, $p < 0.01$) and TextBlob ($\beta = 0.162$, $p < 0.01$). The results of the second regression highlight that when star ratings are included as another predictor, we are able to explain a higher percentage of the variance in the human judge scores (69.6 %). Star ratings make the largest contribution ($\beta = 0.473$, $p < 0.01$), followed by VADER ($\beta = 0.217$, $p < 0.01$), then Google ($\beta = 0.202$, $p < 0.01$) and TextBlob ($\beta = 0.101$, $p < 0.01$). The results of an additional stepwise regression (shown in Table 2) reveal that the addition of each predictor causes a statistically significant increase in R-square, hence, confirming that each tool makes a significant contribution to our ability to predict human judge scores. Therefore, hypothesis 1 is supported.

5.2. Accuracy of SA scores

While the regression results provide evidence that SA tools are adequate predictors of the true sentiment in a review, they do not allow for an assessment of the individual accuracy of these tools, and how these accuracies compare to each other and that of star ratings or combined scores. To be able to do that, we defined tool accuracy as the (absolute) difference between the average human judge score (which ranged from 0 to 100) and the star rating/SA score. Therefore, for each review, we calculated the absolute difference (error) between the true sentiment expressed in that review (as determined by the human judges) and the SA scores generated by each tool or the star rating provided by the original author of the review (i.e., *abs [judge score – tool score]*). In other words, to calculate the score error, we looked at the difference between the tool score (e.g., the score generated by VADER) for each review, and the human judge score for the same review. The absolute value of that difference in scores represents the error of each tool for predicting the "true" score (as determined by the human judges) for each review. For example, the score error of a tool for a review that has a human judge score of 0.65 and a tool score of 0.50 is: *abs (0.65 – 0.50) = 0.15*. We call this variable "score error" in the subsequent analyses.

Table 3 depicts the average score error for each of the three SA tools as well as star ratings. Overall, all tools display a high level of accuracy in reflecting the true sentiment of reviews. Both Google and TextBlob are more accurate (error rates: 14.62 % and 16.27 %, respectively) compared to the star ratings (17.91 %).

To test hypothesis 2, in addition to the scores independently generated by the three SA tools and star ratings, we examined whether meaningful combinations of these scores can improve accuracy. To do so, we created four additional aggregate scores. The first score is a simple heuristic that represents the average of the scores from the three SA tools (Averaged-Score, or AVS for short). Since we expect prediction errors from all these tools (and star ratings) to be largely independent of each other, this simplistic combination is considered worthy of examination. Similarly, AVS+, represents the average of the scores from the three SA tools and star ratings.

Two additional aggregate scores were computed using the unstandardized regression coefficients when predicting the judge scores. The scores were calculated using the coefficients from Regression #1 and Regression #2 depicted in Table 1. To calculate the first score (RS), the three individual SA scores were weighted according to their unstandardized coefficients in Regression #1. Since by the very nature of regression analysis, these coefficients were estimated to minimize the score error, the first regression score (RS), which represents the predicted values using the three SA tools, should be a benchmark for SA scores (and their various combinations including AVS). To calculate the second regression score (RS+), the three SA scores and star ratings were weighted according to their unstandardized coefficients in Regression # 2. Hence, the second regression score, which represents the predicted values using the three SA tools as well as the star ratings, should be a benchmark for the combination of SA scores and star ratings (including AVS+). The results in Table 3 highlight that aggregate scores have higher accuracy (lower average error) compared to any of the four individual scores.

To test whether these differences are significant, while accounting for the effects of contextual factors, we performed an ANOVA where review length (short, medium, long) and product/service category (search, experience, credence) served as two factors, in addition to the main factor representing the tool scores (star ratings, Google scores, VADER scores, TextBlob scores, AVS, AVS+, RS, and RS+). Score error is the dependent variable in the analysis. This resulted in an effective sample size of 7200 (8 scores for each review), where each data point represents the score error for a specific review.

The results of the ANOVA are depicted in Table 4. As the results highlight, the type of tool has a statistically significant main effect on score accuracy (i.e. score error) ($F = 123.5$, $p < 0.01$). The results from a post-hoc analysis (depicted in Fig. 2; summarized in Table 5) reveal that the differences in accuracy between any of the four combined scores and any of the individual tool scores (i.e., Stars, Google, VADER, TextBlob) are all statistically significant. This lends initial support for hypothesis 2. By default, RS+ (computed using the coefficients from Regression #2) displays the best accuracy (9.8 %). This is followed by AVS+ (which is represents the linear average of the three SA scores and star ratings) with an average error rate of (10.5 %). Given the significant interactions between tool and review length as well as between tool and product/service type, before hypothesis H2 can be fully confirmed, the same effects need to be observed for different review length and product categories (see Fig. 3a and b).

The post-hoc analysis further reveals that the differences in accuracy between RS+ and AVS+ is not statistically significant. The other two combined scores, which do not incorporate star ratings, namely RS and AVS perform slightly worse (12.0 % and 12.5 %, respectively).

[1] Our dataset including the reviews and corresponding sentiment labels is available upon request from the authors.

Table 1
Regression results.

	Regression #1					Regression #2				
	B	Std. Er.	Beta	t	Sig.	B	Std. Er.	Beta	t	Sig.
(Constant)	−0.010	0.026		−0.379	0.704	0.056	0.022		2.554	0.011
Google	0.391	0.027	0.401	14.594	< 0.001	0.197	0.025	0.202	7.942	< 0.001
VADER	0.236	0.021	0.330	11.467	< 0.001	0.155	0.018	0.217	8.689	< 0.001
TextBlob	0.320	0.056	0.162	5.743	< 0.001	0.198	0.047	0.101	4.176	< 0.001
Star Rating						0.314	0.017	0.473	18.989	< 0.001
	Adjusted R^2 = 0.573					Adjusted R^2 = 0.696				

Table 2
Results of stepwise regression.

Model	Independent variables	Regression coefficient (t-value)	Δ R^2 (total R^2)
1. Base Model	Star Rating	0.768 (t = 35.882)	**0.589**
2. Model 2	Star Rating	0.590 (t = 25.484)	**+0.074**
	VADER	0.325 (t = 14.019)	(0.662)
3. Model 3	Star Rating	0.487 (t = 19.553)	**+0.028**
	VADER	0.255 (t = 10.875)	(0.690)
	Google	0.226 (t = 9.013)	
4. Full Model	Star Rating	0.473 (t = 18.989)	**+0.006**
	VADER	0.217 (t = 8.689)	(0.696)
	Google	0.202 (t = 7.942)	
	TextBlob	0.101 (t = 4.176)	

Note. Bold numbers indicate a significant R^2 change.

Table 3
Average Tool Accuracy.

Tool	Score Error	
	Mean Error (%)	Std. Deviation
Star Ratings	17.91	14.75
Google	14.62	12.80
VADER	22.91	17.18
TextBlob	16.27	11.32
AVS (average of three SA scores)	12.52	10.70
AVS+ (average of three SA scores and stars)	10.55	8.84
RS (based on Regression#1)	12.01	9.51
RS+ (based on Regression#2)	9.83	8.41

Table 4
ANOVA results.

Effect	df	F	Sig.
Tool	7	123.481	< 0.001
Product Category	2	21.465	< 0.001
Review Length	2	69.122	< 0.001
Tool * Product Category	14	2.492	0.002
Tool * Review Length	14	7.518	< 0.001
Product Category * Review Length	4	13.676	< 0.001
Tool * Product Category * Review Length	28	0.423	0.997

Post-hoc analysis indicates that while the difference in accuracy between RS and AVS is not statistically significant, the differences in accuracy between AVS+ and AVS, and between RS+ and RS are statistically significant. Finally, the analysis further indicates that the differences between Google (14.6 %) and TextBlob (16.3 %), and between TextBlob and star ratings (17.9 %) are all not statistically significant.

5.3. The effects of contextual factors

As seen in Fig. 2, the average of the tool scores with (AVS+) or without (AVS) star ratings are not significantly different from the benchmark scores RS and RS+. Therefore, for the sake of parsimony, in the following analyses we only focuses on a subset of the aggregate scores, namely AVS and AVS+.

The results of three ANOVAs, each focusing on a single product category, reveal a number of interesting effects. In the case of *search products*, tool has a significant main effect. Post-hoc analysis reveals that the accuracy of either one of the two combined scores (i.e., AVS and AVS+) is higher and statistically different from TextBlob, Stars or VADER. However, while the accuracy of AVS + is significantly better than Google, the accuracy of AVS is not (see Fig. 4). Therefore, hypothesis H2 is only partially supported in the context of search product reviews.

In the case of *experience products*, tool has a significant main effect as well. Post-hoc analysis reveals that the accuracy of star ratings is better for reviews of this category, and although it remains to be lower than and statistically different from AVS+, it is not statistically different from AVS. Similarly, Google's accuracy is no different than that of AVS (or Stars). On the other hand, both combined scores exhibit higher accuracies than TextBlob and VADER, with the accuracy of VADER remaining to be significantly the worst (see Fig. 5). Hence, hypothesis H2 is only partially supported in the context of experience product reviews.

In the case of *credence products*, tool has a significant main effect. Post-hoc analysis reveals that Google's accuracy is not significantly different from either of the two combined scores. On the other hand, both combined scores exhibit accuracies that are higher and statistically distinguishable from TextBlob, Stars or VADER. Furthermore, for credence products, the accuracy of star ratings is lower and different from Google and TextBlob, but is not different from VADER (see Fig. 6). Hence, hypothesis H2 is only partially supported in the context of credence product reviews.

A similar analysis of the effects of tool and product across length categories indicates that in the case of *short reviews*, tool has a significant main effect. Post-hoc analysis (Fig. 7) reveals that the accuracy of either combined score is superior to, and statistically distinguishable from, any of the individual tools. The results also reveal that although the difference in accuracy between TextBlob and Google is not statistically significant, the former outperforms the latter, while star ratings exhibit the worst accuracy (while not statistically different from that of VADER). This lends support to hypothesis H2 in the context of short reviews.

In the case of *medium length* reviews, tool has a significant main effect. The difference in accuracy between the two combined tool scores is not statistically significant. The results suggest that while Google's accuracy is statistically no different than AVS, it is statistically different from that of AVS+. The results further reveal that the accuracy of either combined scores is higher and statistically different from TextBlob, Stars or VADER, while TextBlob's accuracy is no different than Google or star ratings, and VADER remains to exhibit the worst accuracy (see Fig. 8). Hence, hypothesis H2 is only partially supported in the context of medium length reviews.

Finally, in the case of *long reviews*, tool has a statistically significant main effect. The accuracies of the two combined scores and Google are

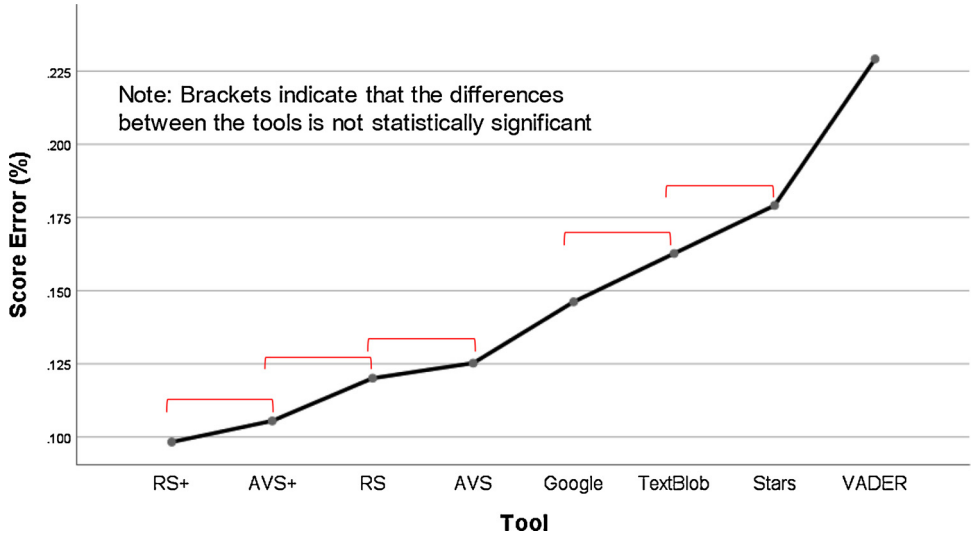

Fig. 2. Tool Main Effects.

Table 5
Pairwise comparisons of tools.

	AVS	AVS+	Google	Stars	TextBlob	RS	RS+	VADER
AVS		0.020 (0.009)	−0.021 (0.004)	−0.054 (< 0.001)	−0.037 (< 0.001)	0.005 (0.984)	0.027 (< 0.001)	−0.104 (< 0.001)
AVS+	−0.020 (0.009)		−0.041 (< 0.001)	−0.074 (< 0.001)	−0.057 (< 0.001)	−0.015 (0.144)	0.007 (0.899)	−0.124 (< 0.001)
Google	0.021 (0.004)	0.041 (< 0.001)		−0.033 (< 0.001)	−0.017 (0.059)	0.026 (< 0.001)	0.048 (< 0.001)	−0.083 (< 0.001)
Stars	0.054 (< 0.001)	0.074 (< 0.001)	0.033 (< 0.001)		0.016 (0.061)	0.059 (< 0.001)	0.081 (< 0.001)	−0.050 (< 0.001)
TextBlob	0.037 (< 0.001)	0.057 (< 0.001)	0.017 (0.059)	−0.016 (0.061)		0.043 (< 0.001)	0.064 (< 0.001)	−0.066 (< 0.001)
RS	−0.005 (0.984)	0.015 (0.144)	−0.026 (< 0.001)	−0.059 (< 0.001)	−0.043 (< 0.001)		0.022 (0.002)	−0.109 (< 0.001)
RS+	−0.027 (< 0.001)	−0.007 (0.899)	−0.048 (< 0.001)	−0.081 (< 0.001)	−0.064 (< 0.001)	−0.022 (0.002)		−0.131 (< 0.001)
VADER	0.104 (< 0.001)	0.124 (< 0.001)	0.083 (< 0.001)	0.050 (< 0.001)	0.066 (< 0.001)	0.109 (< 0.001)	0.131 (< 0.001)	

Note: Number outside parentheses represents the mean difference of score error. Number inside the parenthesis represents the p-value for the mean difference (i.e., whether the difference is statistically significant or not).

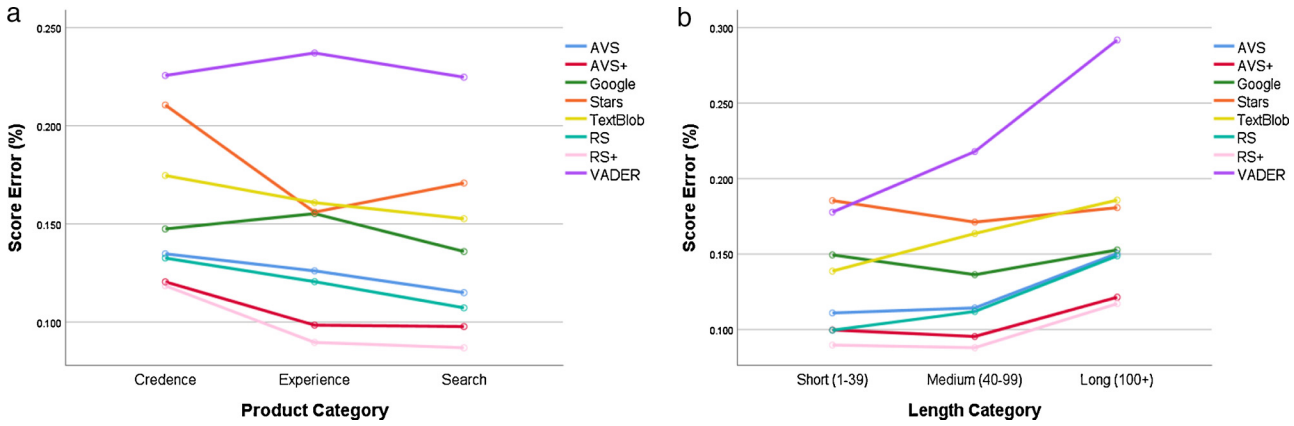

Fig. 3. a. Tool and Product Category. b. Tool and Review Length.

not statistically different. Interestingly, the accuracy of AVS is also not statistically distinguishable from star ratings, while the difference in accuracy between Stars and TextBlob is also not statistically significant. Both TextBlob and VADER exhibit accuracies that are distinguishable from and lower than the two combined scores, while VADER's accuracy is significantly worse than the other tools (see Fig. 9). Therefore, only partial support is obtained for H2 in the context of long reviews.

Table 6 summarizes the results for testing hypothesis H2. Specifically, it indicates whether a proposed difference in accuracy between the combined scores and an individual score are true for both AVS+ and AVS, only for one of them, or neither. For instance, as reported earlier, the difference in accuracy between Stars and the combined scores is statistically significant for both AVS+ and AVS in the context of search and credence products and short and medium reviews, but statistically significant when comparing only Stars and AVS+ in the context of experience products and long reviews.

5.4. Tool type analysis

To test hypotheses 3 and 4, we performed a number of additional ANOVAs. Each ANOVA focused on one specific tool and included length and product category as the two factors predicting the tool's accuracy. The results are presented in Table 7, where each row represents the results from the ANOVAs and post-hoc analysis for an individual tool.

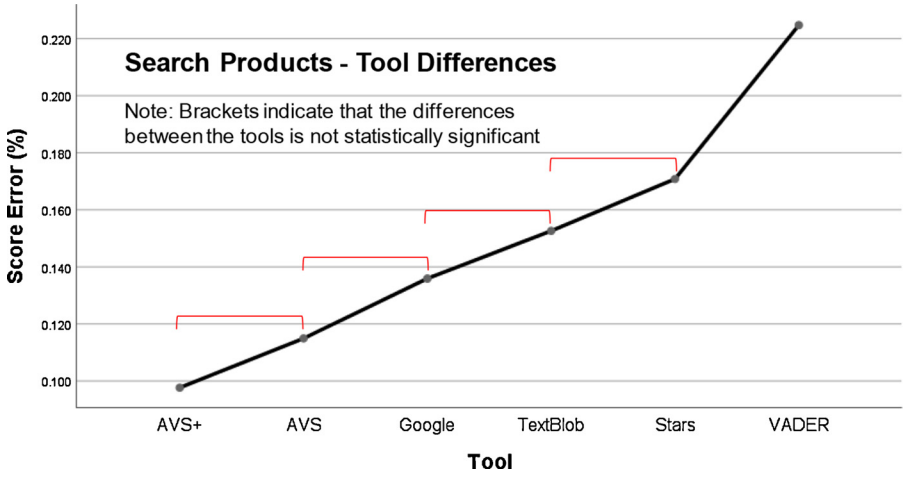

Fig. 4. Tool Effects – Search Products.

Specifically, the second column (from the left) in the table indicates whether the main effect of length category is significant in the ANOVA for each tool. For example, as the results in the table show, length category exerts a main effect on the accuracy of VADER and TextBlob, where the relative accuracies of either tool across the three length categories (i.e., short vs. medium; short vs. long; medium vs. long) are statistically different. The three subsequent columns indicate whether the difference in accuracy between short and medium, short and long, and medium and long (respectively) are significant. The subsequent four columns present similar results but for the product category main effect, and the differences in the accuracy of each tool across the different product categories. The last column in the table indicates whether the interaction of product category and length category is significant for each specific tool. For example, for star ratings the interaction between product and length categories is significant indicating that the effects of these two variables extend beyond their own individual effects represented in their main effects.

Hypothesis 3a states that review length moderates the accuracy of lexicon-based SA tools, such that the accuracy is higher for short reviews then medium reviews then long reviews. As the results in Table 7 indicate, in the case of VADER, the differences in accuracy between short and medium, short and long, and medium and long reviews are all statistically significant, with short reviews exhibiting the highest accuracy (score error = 17.8 %), followed by medium (score error = 21.8 %), and then long (score error = 29.2 %). Hence, H3a is supported. However, since the results indicate that for VADER, the differences in accuracy between the various product categories are not significant, no support can be obtained for H4a (which proposes that product type moderates the accuracy of lexicon-based tools).

Hypothesis 3b proposes that review length moderates the accuracy of hybrid (ML and lexicon based) SA tools, such that the accuracy is higher for short reviews then medium reviews then long reviews. As shown in Table 7, the accuracy of TextBlob significantly changes across length categories, with short reviews exhibiting the highest accuracy (score error = 13.9 %), followed by medium (score error = 16.4 %), and then long (score error = 18.6 %). Hence, H3b is supported. However, the results also reveal that the accuracy of TextBlob is only significantly different between search and credence products (score error = 15.3 % vs. 17.5 %), but is not statistically different between search and experience (score error = 15.3 % vs. 16.1 %), or experience and credence products (score error = 16.1 % vs. 17.5 %). Hence, only partial support is obtained for H4b (which proposes that product type moderates the accuracy of hybrid SA tools).

In contrast, the accuracies of ML-based tools and star ratings are not impacted by review length. As the results in Table 7 show, the differences in accuracy across the length categories are not statistically significant for either Google or star ratings. Similarly, the accuracy of ML-based tools and star ratings are not impacted by product type. While the results in Table 7 suggest that the accuracy of Google is not statistically different across product types, the accuracy of star ratings changes between search and credence (score error = 17.1 % vs. 21.1 %), and experience and credence products (score error = 15.6 % vs. 21.1 %), while it is not statistically different between search and experience products.

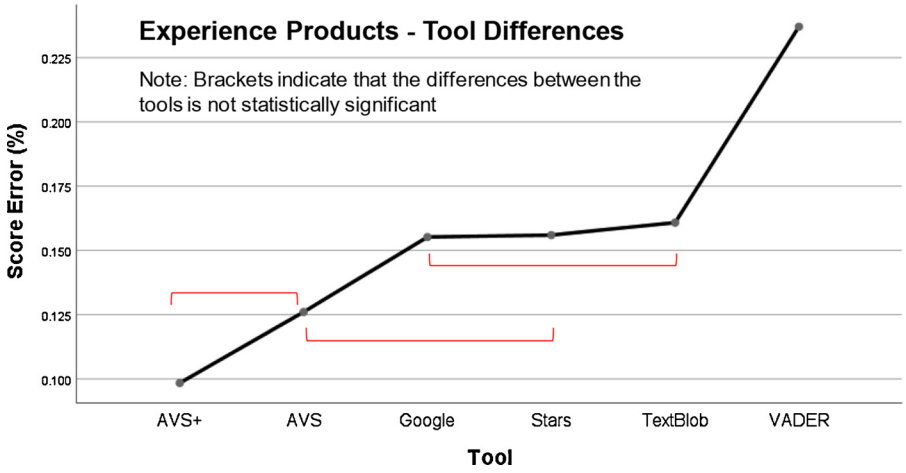

Fig. 5. Tool Effects – Experience Products.

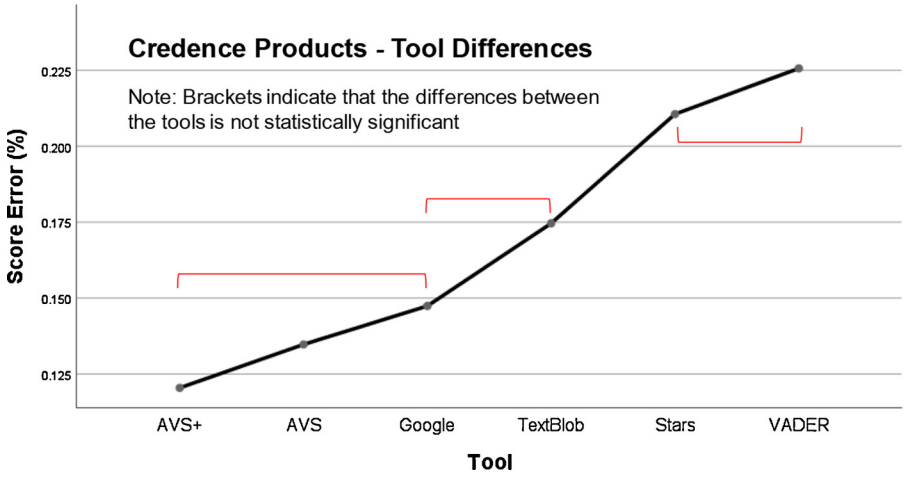

Fig. 6. Tool Effects – Credence Products.

6. Discussion

6.1. Discussion of the results

Table 8 summarizes the level of support obtained for each of the hypotheses. As the table highlights, with the exception of H4a, we are able to obtain full or partial support for all the hypotheses. Consistent with prior tool-specific (e.g., Moraes et al., 2013) or comparison (e.g., Hasan et al., 2018) studies, the results from this study, validate the proposition that SA tools have the potential to accurately detect the true sentiment expressed in a consumer review. Hence, SA scores can be used as complements to, or substitutes for, star ratings (Valdivia et al., 2017). While the accuracy of different tools can vary significantly, our results suggest that the addition of any tool can contribute significantly to enhancing the accuracy of detecting a review's sentiment (Hypothesis 1).

Interestingly, aggregate scores that are based on a combination of multiple SA tools, especially when accounting for star ratings, exhibit accuracies that are higher than any of the individual sentiment scores (Hypothesis 2). This highlights the potential for synergy between different SA tools, and the persistent importance of star ratings as a factor that contributes to increased accuracy. This is a striking finding, especially when considering that our aggregate score is a simple average of the SA scores from the three tools with and without the reviewer-assigned star ratings. While prior research has recognized the potential of SA tools that are based on different approaches to generate uncorrelated and largely divergent scores (e.g., Liu, 2012), and the potential of these tools to act as complements (Appel et al., 2016; Prabowo & Thelwall, 2009), no studies have looked at the potential enhancements that can be obtained by simply combining the outputs of existing tools (especially those that are available off-the-shelf). Instead, the focus of past research has been on modifying the algorithms that form the basis of these tools, in order to create hybrid tools that embody selective choices. In contrast, our results show that combining the outputs of individual tools can be a more fruitful approach than attempting to amalgamate their features. Potentially, the accuracy of combined scores can be further enhanced by considering additional tools, and especially those generating SA scores using unique approaches, even if these individual scores are not especially accurate.

The findings in relation to the relative accuracy of combined scores when examined in the context of specific length or product category length are interesting. While a combined score that accounts for both SA scores and star ratings can always outperform star ratings in any length or product context, the accuracy of a combined score that is solely based on SA scores is not statistically different from star ratings for experience products and long reviews. This highlights that star ratings, even if not always faithful representations of the true sentiments, can substantially enhance the accuracy of combined scores when the reviews concern subjective evaluations of product/services (experience products), or when the reviews are long. This stands in clear contrast to prior comparison studies that attempted to fully substitute sentiment scores for star ratings, even when star ratings are available (e.g., Valdivia et al., 2017).

A similar pattern is observed when the accuracy of combined scores

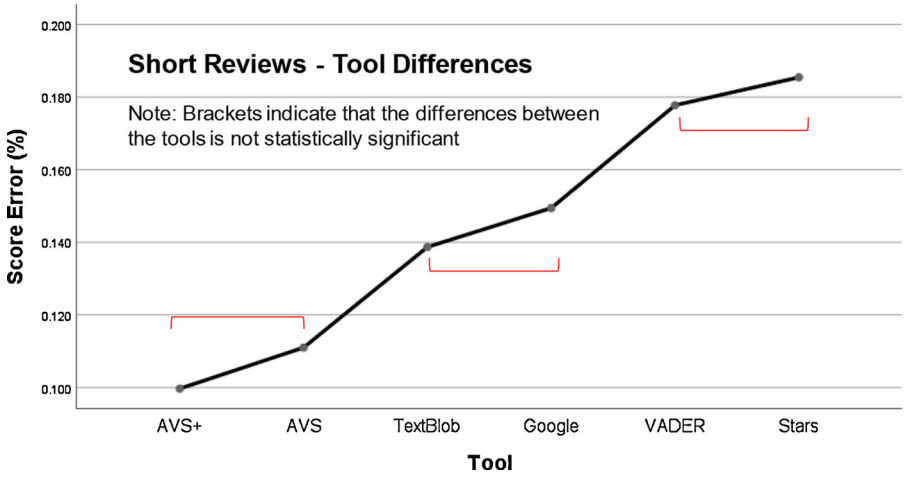

Fig. 7. Tool Effects – Short Reviews.

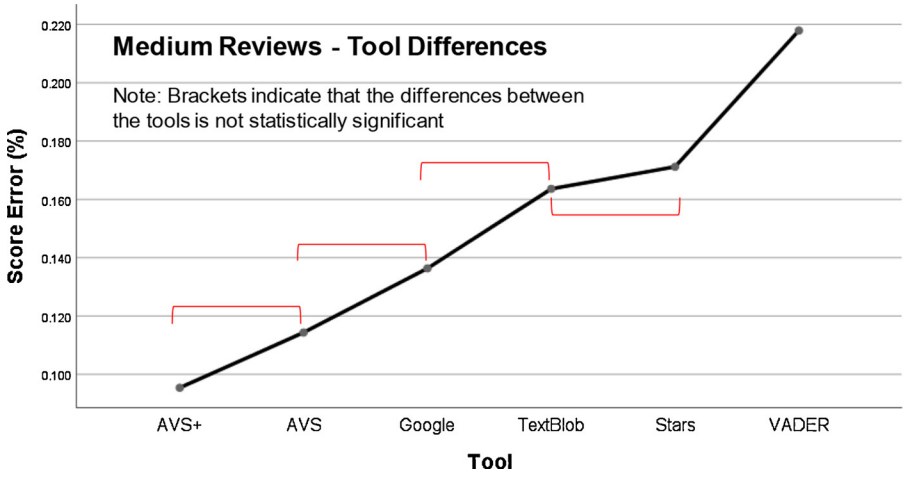

Fig. 8. Tool Effects – Medium Reviews.

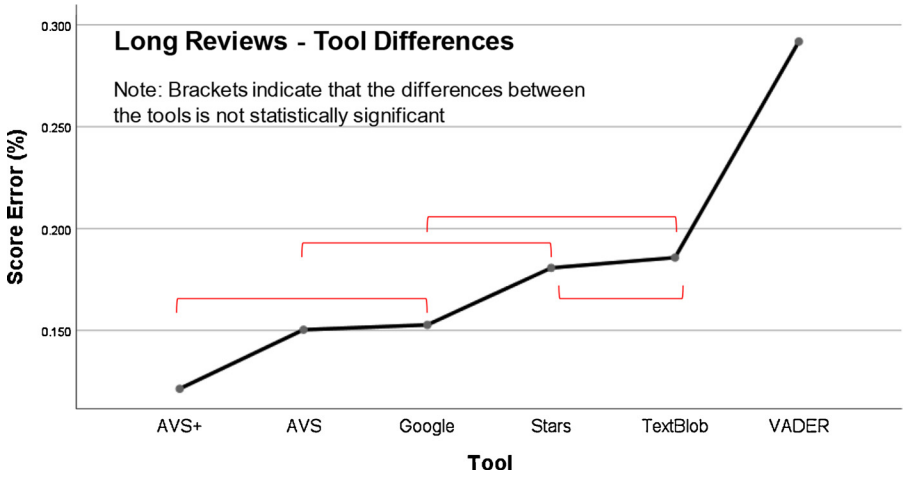

Fig. 9. Tool Effects – Long Reviews.

is compared to the accuracy of ML-based SA scores for different product and length categories. A combined score that accounts for star ratings can outperform ML-based SA scores in most cases, except when the review is long or concerns a credence good. This could be due to the fact that the combined score performs relatively worse in these two contextual categories (probably due to the relatively higher error rates in the lexicon-based scores), while as evidenced by the results, the ML-based tool (i.e., Google) performs consistently across all categories. On the other hand, the accuracy of a combined score that does not account for star ratings is not statistically different from ML-based scores in almost all contexts bar one: short reviews. This could be caused by the relatively high accuracy of the ML-based tool, which is consistent with recent SA studies that show that machine-learning approaches to SA, especially those employing neural networks exhibit high level of accuracy (e.g., Moraes et al., 2013; Ye et al., 2009). This reaffirms our other findings regarding the consistently-observed high accuracy of ML-based scores, especially when compared to a combined score that only accounts for other SA-generated scores (and hence, a score that does not account for the unique variation captured by star ratings). Nonetheless, in relation to these findings, it is important to note that although the difference in accuracy between the ML-based score and the combined score is not statistically significant, combined scores always exhibit higher levels of accuracy.

As suggested by the results, overall, the length of a review can significantly affect the accuracy of SA scores, and so does the type of product/service. This corroborates earlier suggestions that there are qualitative differences between reviews of different lengths or addressing different product categories (Mudambi & Schuff, 2010), which as

Table 6
Summary of results for H2.

Tool	Product category			Review length		
	Search	Experience	Credence	Short	Medium	Long
Star ratings	AVS AVS+	AVS+	AVS AVS+	AVS AVS+	AVS AVS+	AVS+
Lexicon-based (VADER)	AVS AVS+	AVS AVS+	AVS AVS+	AVS AVS+	AVS AVS+	AVS AVS+
ML-based (Google)	AVS+	AVS+	none	AVS AVS+	AVS+	none
Hybrid (TextBlob)	AVS AVS+	AVS AVS+	AVS AVS+	AVS AVS+	AVS AVS+	AVS AVS+

Table 7
Length and product type differences per tool.

	Length Main Effect	Short → Med.	Short → Long	Medium → Long	Product Main Effect	Search → Experience	Search → Credence	Experience → Credence	Product * Length Interaction
Star Ratings	N	N	N	N	Y	N	Y	Y	Y
Google	N	N	N	N	N	N	N	N	N
VADER	Y	Y	Y	Y	N	N	N	N	N
TextBlob	Y	Y	Y	Y	Y	N	Y	N	N

Y = Indicates that the effect or the difference is significant at $p < 0.05$.
N = Indicates the effect or the difference is not significant.

proposed in prior research, has the potential to affect SA accuracy (Agarwal et al., 2011; Liu, 2012). The tool-based results further suggest that the effects of product context and review length vary for each tool (Hypotheses 3 and 4). Specifically, while the accuracy of star ratings and ML-based scores does not appear to change across reviews of different lengths, the accuracy of the scores generated by lexicon-based and hybrid approaches is highest for short reviews, followed by medium ones. This is consistent with our reasoning that because lexicon-based tools or hybrid tools that use lexicons average the sentiments of individual words (Hutto & Gilbert, 2014), the resultant overall scores will likely not account for nuances or contrasting opinions that are often common to longer reviews.

The results concerning the moderating role of product category on sentiment score accuracy are most intriguing. Contrary to our prediction (in H4a), the accuracy of scores generated using a lexicon-based tool did not significantly change across product categories. This reaffirms VADER's status as the gold-standard of lexicon-based SA across varied types of textual content (Ribeiro et al., 2016; Hutto & Gilbert, 2014). A possible explanation for this is that unlike what the theory implies, the relative subjectivity of the reviews in different categories is not highly differential, and pure lexicon-based tools can perform similarly to ML-based tools, and especially those that apply deep learning in this context (Kirilenko et al., 2018; Moraes et al., 2013). The results concerning the effects of product context on score accuracy for SA scores generated by a hybrid tool are more mixed. While the accuracy of scores does not differ between search and experience products, it does differ between search and credence products. This highlights that contextual factors can affect the accuracy of hybrid tools differently than they affect the accuracy of tools that employ one of the approach embedded within these hybrid tools.

6.2. Theoretical implications

This study makes a number of theoretical contributions. First, it affirms the important role of the SA approach on the utility of surrogate tools. As our findings indicate, choices concerning the overall approach adopted by an SA tool (ML, lexicon, hybrid), and subsequent algorithmic and training choices, not only impact their accuracy, but their overall scalability and generalizability to different domains.

Second, the results highlight the importance of considering contextual factors when designing and evaluating the accuracy of SA tools. While in this study, we only focus on two contextual factors chosen based on extant research, there is potential for additional factors such as characteristics of the consumer to play an important role in other types of contexts, e.g., analyzing political sentiment.

Third, findings from this study present clear evidence regarding the significant performance improvement that can be attained from integrating multiple tools. While some past research has experimented with creating hybrid tools, our results show that even the less-than-perfect individual tool, can make significant contributions to the accuracy of a combined score. Such findings not only expand our view of the potential of SA, but also offer general methodological insights on how to design studies that are focused on comparing different technological artifacts, and the importance of considering combinations of the outputs of these artifacts.

A fourth major theoretical contribution of this study is the systematic and experimental approach used to answer the research questions posed. Modeling sentiment accuracy as the ability of the tools to approximate the true sentiment expressed in a review as determined by human judges, eliminates any biases inherent in approaches that use star ratings, which, as past literature and our results suggest, are less than perfect representatives of the true sentiment of a review. Also, employing a full factorial experimental design ensured the control of any confounds, and allowed for a more refined investigation of the effects of contextual factors. Finally, modeling SA scores as continuous variables, allows for the detection of small changes in score accuracy that is difficult to detect when adopting the nominal approach to modeling SA.

6.3. Practical implications

The results suggest some fairly consistent patterns in the performance of various approaches to SA. First and foremost, the results establish that regardless of the context, SA tools can produce scores that

Table 8
Summary of the hypotheses and level of support obtained.

Hypothesis	Level of support obtained
H1: Sentiment scores that are based on different approaches act as significant predictors of the true sentiment of a review.	**Supported**
H2: The accuracy of combined sentiment scores in predicting the true sentiment of a review will be higher than the accuracy of individual scores.	**Partial Support** (see Table 6)
H3a: Review length moderates the accuracy of lexicon-based SA tools such that the accuracy is higher for short reviews then medium reviews then long reviews.	**Supported**
H3b: Review length moderates the accuracy of hybrid SA tools such that the accuracy is higher for short reviews then medium reviews then long reviews.	**Supported**
H4a: Product type moderates the accuracy of lexicon-based SA tools such that the accuracy is the highest for search products then experience products then credence products.	**Not Supported**
H4b: Product type moderates the accuracy of hybrid SA tools such that the accuracy is the highest for search products then experience products then credence products.	**Partial Support** - Supported between search and credence.

more faithfully approximate the true sentiment of a review than star ratings can: at least one SA score outperformed star ratings in each scenario we examined. The suggestion is that integration of sentiment analysis tools into the presentation of consumer reviews should certainly be a consideration in how reviews are presented. Meanwhile, as the regression analysis in Table 2 indicates, star ratings seem to capture a somewhat unique aspect of the variance in the sentiment of a review; therefore, whenever they are available, star ratings have utility especially if they are combined with SA scores. Among the SA tools, the ANN-based Google is the most robust as its performance does not seem to be affected by the contextual factors. Google also consistently outperforms star ratings and the other SA tools (except for short reviews where it is tied with TextBlob). The accuracy of the tools decreases for long reviews, and mostly for experience and credence products, yet the relative performance of the tools is fairly consistent for these categories as well.

Based on the above, we recommend that if a review site or a review portal is to adapt one SA tool, an ANN or deep learning based tool is recommended. Yet, given the superiority of aggregate scores over individual ones, when possible, individual tool scores should be combined with other SA scores and/or star ratings. In fact, the results suggest that even simple average scores perform better than individual scores under various scenarios (and are surprisingly close to the benchmarks). Hence, in practical settings, a simple average of available scores from fundamentally different SA approaches should better reflect the true sentiment of a review.

The results also indicate that the general trend in academia and practice where machine learning based analytics techniques are increasingly preferred over those that require knowledge engineering is justified as VADER, the best lexicon based tool reported in the literature, consistently underperformed the other SA tools that integrate machine learning in their algorithms. VADER was also the only SA tool that consistently underperformed star ratings (except in the case of short reviews where the differences are not statistically different).

The moderation analyses highlight a number of interesting effects. For example, the overall accuracy of SA tools is significantly better for short reviews where the accuracy of star ratings relative to SA tools is lower. This indicates that SA tools are a most valuable substitute for star ratings especially in the case of short reviews. Overall, the results concerning the important moderating role of product context highlight the need to consider contextual factors when designing the algorithms underlying SA tools, or the lexicons that they utilize. It is important that ML-based tools be trained using data sets from different domains and of various lengths. The dictionaries used by lexicon-based tools need to be compiled while considering various contexts. The evident weakness of the lexicon-based approach when analyzing long text points to a potentially inherent deficiency of such tools. While due to their transparent nature, lexicon-based tools will likely continue being used and preferred by many, caution should be exercised when a lexicon-based tool is used with a long piece of text.

6.4. Limitations and future research

As with any research, this study has a number of limitations. The first is a consequence of the numeric transformation we applied to sentiment scores and star ratings in order to be able to compare them to each other and human judge scores. In order to attain the highest possible level of fidelity, we had our judges rate the reviews on a continuous scale of 0–100. While we believe this is a strength of our design, the star ratings are inherently ordinal (1–5), therefore their transformation to a scale out of 100 inevitably created a discrete (scores of 0, 25, 50, 75, and 100) rather than a purely continuous scale such as the scores obtained from the sentiment analysis tools that we selected. Instead of grouping those scores into five categories, we converted star ratings to a scale of 100. This may have reduced the accuracy of our findings that included star ratings. However, when aggregate ratings are presented by review sites, they are typically averages, and hence a similar transformation is performed in practice. While we think a scale of 1–5 is not sufficiently granular, it may be worthwhile for future research to examine, through user studies, what the ideal granularity of these scales should be.

Secondly, although ours was a large and diverse sample, making our comparative analysis more generalizable than what has been previously reported in the literature, the results should still be interpreted with caution especially for those settings that significantly differ from what is included in our analyses. For example, doctor, lawyer or university reviews are hard to categorize into a specific product/service context, and may inherently contain different types of sentiments that may be easier/harder to automatically detect. If one's interest is in those narrower and more specific contexts, the analysis presented here should be replicated.

Thirdly, although combining scores from tools that represent different approaches to SA produced favorable results, we cannot claim that one can indefinitely consider more tools and combine their scores with significant increase in SA accuracy. As the stepwise regression results depicted in Table 2 show, the relative improvement in predicting true sentiment (represented by the R-square change) follows a diminishing returns pattern with the inclusion of each additional tool. While our approach of selecting one state of the art tool representing each major SA approach demonstrates the enhanced utility of combined scores, the theoretical (based on marginal accuracy) and practical (based on time and effort needed) limitations of this approach should be tested.

As we discussed before, the ultimate goal of presenting sentiment scores (or star ratings) along with corresponding reviews is to help the reader decide what might be a useful review before (s)he reads it. Now that we have determined the relative accuracy of various tools and their combinations, future research should further investigate the utility of SA scores in consumer decision making. Specifically, whether the provision of SA scores improves decision confidence and quality in various contexts, and whether users are willing to adopt these decision aids in their regular consumption of review content should be explored further. In such explorations, a number of factors should be considered. As discussed above, the nature of the scale through which sentiment scores are presented is likely a non-trivial issue that affects user performance and adoption intentions. Another related issue is whether a simple numeric score or a graphic presentation (e.g. a bar displaying the relative distance of the score from a neutral point as well as negative and positive extremes) is the most effective way to present these scores. The use of other visual tools such as color and emoticons should also be explored.

Most sentiment analysis research and state of the art SA tools provide the user with a uni-dimensional score (Chang et al., 2017). While displaying a single score representing the overall sentiment can quickly assist the user in determining whether the review is worthy of reading, presenting SA scores in multiple dimensions each representing a different aspect can be more informative. Particularly, this approach aids the customer to efficiently understand the reasoning behind the negative/positive review, likely without the need to read the review in detail. Hence, a refined score will significantly increase accuracy by allowing the customer to gain an understanding of the review's essential elements, without adding much additional effort.

A more informative SA score based on the different sentiments included in the review could potentially lead to different attitudes and behaviors. This approach is based on research in psychology, which has suggested that there are five major sentiments (happiness, sadness, anger, disgust, and fear), the effects of which are similar regardless of culture (Ekman, 1992). These sentiments can guide human behavior (Wu et al., 2019), even though different sentiments with opposite polarity could have similar consequences. For example, both happiness and anger increase confidence and feeling of power while decreasing sensitivity to risk (Lerner & Tiedens, 2006), and therefore can lead to

similar behavior (e.g., willingness to pay an unusually high price for a product). Some research has examined this issue of emotion recognition from text (termed Emotional Text Mining, Greco & Polli, 2019), where categorization of tweets into "angry", "disgusting", "joyful" and "sad" led to the identification of certain mood patterns and abnormal events according to those patterns (Zhao et al., 2012). Similarly, Wu et al. (2019) used the Ortony-Clore-Collins (OCC) model of emotion to capture the emotions embedded in microblogging data concerning emergency events. Currently, there are very few tools that can capture these sentiments from written text, yet, the exploration of multi-dimensional expression of sentiments and their influence on decision making remains an interesting avenue of research.

7. Conclusion

This research investigated the ability of SA tools to accurately detect the true sentiment expressed in a consumer review. The results of an empirical investigation of multiple SA tools, along with star ratings indicate that SA scores can reflect a review's sentiment with fair levels of accuracy. This confirms their potential to act as substitutes in contexts where star ratings are unavailable. The results of the study further highlight that integrating the output of multiple SA tools, and integrating those with star ratings can significantly enhance accuracy in detecting a review's true sentiment. Hence, we confirm the potential of SA tools to act as complements of star ratings when these are available. Finally, the study sheds light on the important effects of contextual factors of review context and length, which are shown to overall influence the accuracy of SA tools, and moderate the effects of specific tools.

With the growing importance of eWOM in all facets of decision-making, the need for automated and economically viable means of summarizing large amounts of decision-relevant data is essential. This study advocates for the potential of SA tools as a means to alleviate this big data problem, and confirms this potential through a controlled empirical investigation.

CRediT authorship contribution statement

Sameh Al-Natour: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. **Ozgur Turetken:** Conceptualization, Methodology, Resources, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition, Project administration.

Acknowledgments

This research was supported by a joint research grant from Ryerson University and Hong Kong Polytechnic University. We would like to thank Amirkiarash Kiani for his help with collecting some of the data used in this study.

Appendix A. Additional analyses

Beyond the analyses required for hypotheses testing, we examined the other effects that resulted from our tests. The ANOVA results in Table 4 reveal that product category has a statistically significant main effect ($F = 21.5$, $p < 0.01$), with the score error being lower for search products (12.9 %), followed by experience (13.7 %) and credence (15.3 %). Post-hoc analysis indicates that while the difference in accuracy between search and experience is not statistically significant (see Fig. A1), the difference between credence and either of the other two types of product/service category is significant. Review length also has a statistically significant main effect ($F = 69.1$, $p < 0.01$), with the accuracy being higher for short (12.5 %), followed by medium (13.1 %) and then long (16.3 %) reviews (see Fig. A2). Post-hoc analysis indicates that the difference in accuracy for short and medium reviews is not statistically significant.

In the case of *search products*, length has a significant effect: medium length reviews exhibit the best accuracy followed by short reviews (albeit, the difference is not statistically significant), while the only significant difference exists between medium and long reviews (not diagrammed). In the case of *experience products*, length has a significant effect as well: the difference between short and medium length reviews is not statistically significant (not diagrammed). Finally, in the case of *credence products*, length has a significant effect and the differences between the three length categories are all statistically significant (not diagrammed).

In the case of *short reviews*, there are no differences in terms of accuracy between the three types of product categories (not diagrammed). In the case of *medium length* reviews product category has a statistically significant main effect: the accuracy in experience product contexts is not different than that in credence or search product contexts, while the accuracy for search products is better than credence (not diagrammed). Finally, in the case of *long reviews*, product category has a statistically significant main effect. The difference in accuracy between search and experience products is not statistically significant, with the accuracy being better for search, followed by experience and then credence products (not diagrammed).

Overall, the results reveal that most differences relating to product category exist between credence and search products, while differences due to review length are largely caused by the differences between short and medium reviews on one hand, and long reviews on the other.

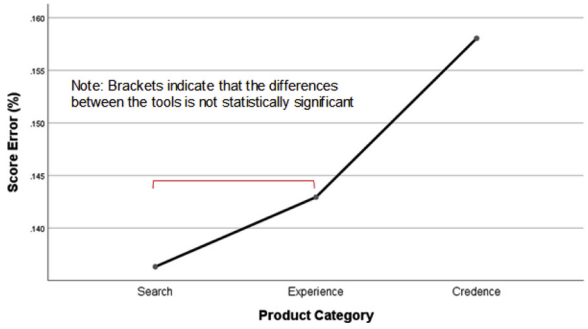

Fig. A1. The Effects of Product Category.

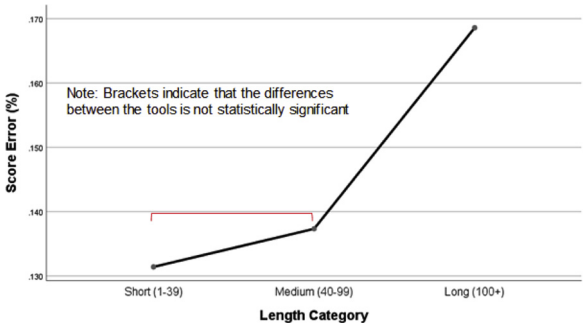

Fig. A2. The Effects of Review Length.

References

Abe, S. (2005). *Support vector machines for pattern classification, Vol. 2*. Springer.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. *Proceedings of the Workshop on Language in Social Media (LSM 2011),* 30–38.

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research, 58*(2), 175–191.

Annett, M., & Kondrak, G. (2008). *A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs*.

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems, 108*, 110–124.

Bulbul, C., Gross, N., Shin, S., & Katz, J. (2014). *When the path to purchase becomes the path to purpose. Think with google*.

Chang, Y.-C., Ku, C.-H., & Chen, C.-H. (2017). Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *International Journal of Information Management*.

Chatterjee, S. (2019). Explaining customer ratings and recommendations by combining qualitative and quantitative user generated contents. *Decision Support Systems, 119*, 14–22.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science, 54*(3), 477–491.

Cheung, C. M., & Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision Support Systems, 53*(1), 218–225.

Chopra, D., Joshi, N., & Mathur, I. (2016). *Mastering natural language processing with python.* Packt Publishing Ltd.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284.

Dang, Y., Zhang, Y., & Chen, H. (2009). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems, 25*(4), 46–53.

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems, 25*(4), 46–53.

De Maeyer, P. (2012). Impact of online consumer reviews on sales and price strategies: A review and directions for future research. *Journal of Product and Brand Management, 21*(2), 132–139.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3-4), 169–200.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89.

Gao, S., Hao, J., & Fu, Y. (2015). *The Application and Comparison of Web Services for Sentiment Analysis in Tourism*.

Ghasemaghaei, M., Eslami, S. P., Deal, K., & Hassanein, K. (2018). Reviews' length and sentiment as correlates of online reviews' ratings. *Internet Research, 28*(3), 544–563.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications, 69*, 214–224.

Greco, F., & Polli, A. (2019). Emotional text mining: Customer profiling in brand management. *International Journal of Information Management*.

Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin, 118*(3), 430.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications, 23*(1), 11.

Hu, Y.-H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management, 36*(6), 929–944.

Hutto, C. J., & Gilbert, E. (2014). *Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social media Text*.

Jeong, B., Yoon, J., & Lee, J.-M. (2017). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management, 48*, 280–290.

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. *Journal of Travel Research, 57*(8), 1012–1025.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Kordzadeh, N. (2019). Investigating bias in the online physician reviews published on healthcare organizations' websites. *Decision Support Systems, 118*, 70–82.

Lak, P., & Turetken, O. (2014). Star ratings versus sentiment analysis–a comparison of explicit and implicit measures of opinions. *2014 47th Hawaii International Conference on System Sciences* (pp. 796–805).

Lerner, J. S., & Tiedens, L. Z. (2006). Portrait of the angry decision maker: How appraisal tendencies shape anger's influence on cognition. *Journal of Behavioral Decision Making, 19*(2), 115–137.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, 2*(2010), 627–666.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1–167.

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M., & Muhammad, K. (2020). A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management, 50*, 432–451.

Mills, J. E., & Law, R. (2004). *Handbook of consumer behavior, tourism, and the internet.* Psychology Press.

Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*(2), 621–633.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon. com. *MIS Quarterly, 34*(1), 185–200.

Nasukawa, T., & Yi, J. (2003). *Sentiment Analysis: Capturing Favorability Using Natural Language Processing*.

Nelson, P. (1970). Information and consumer behavior. *The Journal of Political Economy, 78*(2), 311–329.

Nielsen, M. A. (2015). *Neural networks and deep learning, Vol. 2018*. San Francisco, CA, USA: Determination press.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 115–124).

Pavlou, P. A., & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research, 17*(4), 392–414.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge University Press.

Philander, K., & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management, 55*(2016), 16–24.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information.* Oxford University Press.

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2016). *Sentiment analysis in social networks.* Morgan Kaufmann.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, 3*(2), 143–157.

Pradhan, V. M., Vala, J., & Balani, P. (2016). A survey on Sentiment Analysis Algorithms for opinion mining. *International Journal of Computer Applications, 133*(9), 7–11.

Qiu, L., Pang, J., & Lim, K. H. (2012). Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence. *Decision Support Systems, 54*(1), 631–643.

Ragini, J. R., Anand, P. R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management, 42*, 13–24.

Rathore, A. K., & Ilavarasan, P. V. (2020). Pre-and post-launch emotions in new product development: Insights from twitter analytics of three products. *International Journal of Information Management, 50*, 111–127.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science, 5*(1), 23.

Roussinov, D., & Turetken, O. (2009). Exploring models for semantic category verification. *Information Systems, 34*(8), 753–765.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management, 52*(1), 5–19.

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent

trends and future directions. *Journal of Travel & Tourism Marketing, 32*(5), 608–621.
Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458–464.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307.
Thavasimani, P., & Missier, P. (2016). *Facilitating Reproducible Research by Investigating Computational Metadata*.
Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment analysis in tripadvisor. *IEEE Intelligent Systems, 32*(4), 72–77.
Vijayarani, S., & Janani, R. (2016). Text mining: Open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII), 3*(1), 37–47.
Wang, Y., Yuan, J., & Luo, J. (2015). *America Tweets China: A Fine-Grained Analysis of the State and Individual Characteristics Regarding Attitudes Towards China*.
Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* 347–354.
Wu, P., Li, X., Shen, S., & He, D. (2019). Social media opinion summarization using emotion cognition and convolutional neural networks. *International Journal of Information Management*.
Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527–6535.
Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications, 38*(6), 7674–7682.
Zhao, J., Dong, L., Wu, J., & Xu, K. (2012). *Moodlens: An Emoticon-Based Sentiment Analysis System for Chinese Tweets*.
Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(2), 133–148.