Almansour *et al. Future Business Journal* 2022, **8**(1):3
https://doi.org/10.1186/s43093-022-00114-y

Future Business Journal

REVIEW

Open Access

Text-rating review discrepancy (TRRD): an integrative review and implications for research

Amal Almansour*, Reem Alotaibi and Hajar Alharbi

Abstract

The large number of online product and service review websites has created a substantial information resource for both individuals and businesses. Researching the abundance of text reviews can be a daunting task for both customers and business owners; however, rating scores are a concise form of evaluation. Traditionally, it is assumed that user sentiments, which are expressed in the text reviews, should correlate highly with their score ratings. To better understand this relationship, this study aims to determine whether text reviews are always consistent with the combined numeric evaluations. This paper reviews the relevant literature and discusses the methodologies used to analyse reviews, with suggestions of possible future research directions. From surveying the literature, it is concluded that the quality of the rating scores used for sentiment analysis models is questionable as it might not reflect the sentiment of the associated reviews texts. Therefore, it is suggested considering both types of sources, reviews' texts and scores in developing Online Consumer Reviews (OCRs) solution models. In addition, quantifying the relationship degree between the text reviews and the scores might be used as an instrument to understand the quality of rating scores, hence its usefulness as labels for building OCRs solution models.

Keywords: Sentiment analysis, Text review, Rating score, Correlation

Introduction

With advancements in and rapid expansion of Web 2.0 innovations, more and more people are using blogs, forums, Online Consumer Reviews (OCRs), and online bulletin boards to comment on their personal experiences. Online Consumer Review (OCR) platforms present great opportunities to share customer viewpoints, preferences, and experiences on a broad selection of services and products. Therefore, the resulting agglomeration of Online Consumer Reviews (OCRs) represents a vital information source that consumers can access when selecting a product or service. Gartner research (www.gartner.com) reported in the article "The Future of the Social Customer", 2012 that 40% of consumers use social media as a search tool, 77% check online reviews, and 75% of consumers feel online reviews are more trustworthy than personal recommendations [30]. Additionally, 81% of online consumers indicated that they received helpful information and advice from these reviews. Therefore, online reviews are not only read but also trusted [6]; this is supported as well by another online consumer survey done by Nielsen Global (www.nielsen.com), 2012 [12], in which 70% of respondents trusted the online reviews posted by strangers. In addition to consumer use, OCRs can also be used by commercial enterprises as an openly accessible source of valuable information to better understand preferences and perceptions of consumers. In fact, enterprises can analyse consumer feedback to create effective new strategies for product design [65]. Some companies, such as Sysomos (www.sysomos.com), Radian6 (www.radian6.com), or Bazaarvoice (www.bazaarvoice.com), support listening and monitoring tools on

*Correspondence: aalmansour@kau.edu.sa
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

the web that offer real-time intelligence about the reputations of their customers' products or services [23].

Traditionally, review texts were hard to collect en masse in the "non-connected" world. Online Consumer Reviews (OCRs) are frequently provided by online review websites in a free-text format, such as Yelp and Amazon. While this comprehensive source of information can help individuals and businesses make better decisions, consumers are faced with the daunting task of locating and reading multiple potentially relevant text reviews, which can lead to an information-overload problem. Consequently, there is a crucial need to mine the available valuable data from reviews to understand user preferences and make accurate predictions and recommendations. To simplify this task and make it more time efficient, certain review websites provide a score averaging the ratings of reviews, in addition to text reviews, as shown in Fig. 1.

The most commonly used scheme for visually displaying average review ratings is the five-star scoring system. Since these scores are only computable using numerical ratings, text reviews should be converted into numerical values or star ratings. There are two ways to do this: (1) asking customers to express their opinions on products and services using star ratings or (2) calculate the overall ratings of the text reviews through the use of sentiment prediction techniques. Traditionally, it is assumed that user sentiments, which are expressed in the text reviews, should correlate highly with their score ratings [16, 48]. However, there can be a discrepancy between the text sentiment and the rating, which indicates a non-valuable data source for research studies utilizing review texts or rating scores in their solution models. In addition, some research studies considered only rating scores or review texts, assuming that they were correlated with each other; therefore, these studies might have failed to satisfy their research objectives [35, 47]. This problem has been overlooked in the literature, even though this point helps in identifying the validity of the review to be used for other research studies [3, 66, 67].

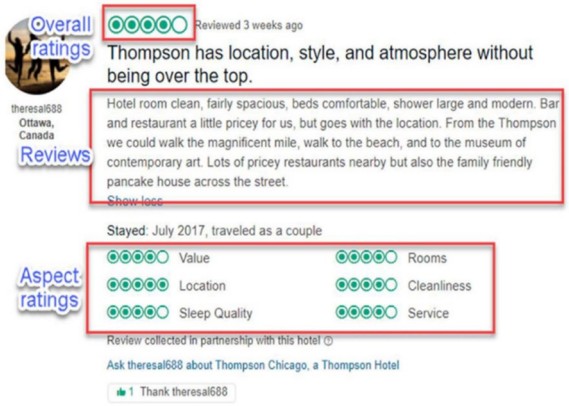

Fig. 1 An example of a text review and corresponding ratings [11]

This paper introduces the concept Text-Rating Review Discrepancy (TRRD), which is defined by the inconsistency between text reviews and the rating reviews of a product or service. In addition, this paper provides a background foundation of text reviews and rating, hence demonstrating the need for using text reviews in addition to rating scores for building Online Consumer Reviews (OCRs) solution models. Therefore, this paper presents a literature review to investigate Text-Rating Review Discrepancy (TRRD) in the evaluation of text reviews and rating scores in detail. Research papers that utilize machine learning algorithms and natural language processing techniques for opinion mining and sentiment analysis evaluation have been considered. Opinion mining and sentiment analysis are important research topics that identify the underlying sentiments in text reviews. Since identifying the accurate text sentiments of consumers' reviews of products and services is important for customers, business owners, and product manufacturers, studying the correlation between online text reviews and ratings is crucial to enhance the correctness of sentiment analysis systems.

This paper examines the following research questions: Are there any discrepancies between text reviews and star ratings? If not, are star ratings considered to be good representatives for text reviews? Will the prediction and recommendation accuracy results improve if the correlation between the text reviews and the associated rating scores are first checked? Are available ratings considered as "gold standard," a high-level representation of ground truth, hence a measurement system whose outputs are known to be accurate and trustable? Ground truth can have wrong labels, it is a measurement, and there can have errors due to human or used machines errors. It should be noted that any trained machine learning model will be limited by the quality of the ground truth used to train and test it. The main contributions of the current study are as follows:

- Discussing the Online Consumer Reviews (OCRs) existing literature and their applications,
- Evaluating numeric rating labels used for building Online Consumer Reviews (OCRs) solution models and showing the necessity of using both text reviews and numeric ratings for building Online Consumer Reviews (OCRs) solution models.

The next following section provides an overview of the importance of studying Text-Rating Review Discrepancy (TRRD) and why we need to use both text reviews and

numeric ratings for building Online Consumer Reviews (OCRs) solution models.

Applications benefit from studying text-rating review discrepancy (TRRD)

Text reviews are a very important source of information for potential consumers before deciding to purchase a product. Consequently, sentiment analysis has a significant impact on products and companies. Many studies used text review to analyse feature specification and customer preferences, assuming the text reviews are consistent with the ratings, which are only general indications of the sentiment.

To deduce information for product creation and product feature selection for business benefits, Xiang et al. [64] used consumer reviews to ascertain what customers want from varying types of hotel. They used text analytics to achieve this task and understand customer preferences. As an example, companies can examine comments left online to understand consumers' feelings or perceptions of a movie and, consequently, predict consumers' interests [52], or use consumer reviews on products to the same end [32, 50]. Also, a literary contribution by Xiao et al. [65] adds to the literature with a preference-measurement model created from consumers' reviews. Textual analysis and the use of the results from this mathematical model aid in understanding consumers' preferences by crowdsourcing from lists of consumers' online reviews. Subsequently the feedback could then be utilized for such things as product redesign.

In their research, Li et al. [30] proposed a social intelligence mechanism that could extract and consolidate reviews given using social media and gain critical insights into new product or service features to assist the decision-making process for the development of new products or services by analysing the reviewers' opinions, authority, and understanding knowledge as well as sentiment towards targeted products. Khalid et al. [27] highlighted some of the issues raised by consumers for mobile-app reviews (e.g. additional cost, practicality, compatibility problems, crashing). They highlighted a statistical depiction of some of the consumer reviews from the Apple App Store and Google Play. In 2014, Vu et al. [62] proposed a keyword-based framework in gathering and getting consumer reviews from the app stores by taking out, evaluating, and categorizing keywords based on semantic resemblance. Additionally, they created an image tool that showed the occurrence of these keywords over a certain period of time and accounted for any suspicious patterns. Park et al. [42] fashioned an app, AppLDA, to be used on app narratives and consumer reviews as a subject model. Using this method, an app developer can examine reviews as well as establish what are seen as essential app features. An automatic system of categorizing customer reviews in regard to programmed classification was offered by Panichella et al. [41]. The system was designed to support the software upholding and requirement progression.

Gu and Kim [21] recommended the use of SUR-Miner to summarize and categorize reviews. They evaluated SUR-Miner on 17 Google Play apps such as Swiftkey, Camera360, WeChat, and Templerun2. They randomly selected and assessed 2000 sentences from text reviews. From different points of view, Mcilroy et al. [36] analysed this study's glitches by looking at the developers of the top apps such as Apple and Google. They observed that there were optimal results, which came from developers replying to the reviews—consumers. This has a positive effect on the reviews as the average ratings has increased by 36.87% and the median rating by 20%.

All discussed research papers above are offering Online Consumer Reviews (OCRs) solution models which use either text review, associated rating numeric or both to build their models. We advise to investigate Text-Rating Review Discrepancy (TRRD) to ensure the validity and correctness of the built solution model. Thus, this paper discusses this topic in detail and is laid out as follows: "Background" section presents background knowledge in the domain of text reviews and ratings, then "Methods" section presents the survey methodology. "Related work" section reviews work on Online Consumer Reviews (OCRs) models and applications and introduces the relevant studies that consider the discrepancies between text reviews and ratings. This study first identifies research that focuses only on one source of information—either text reviews or rating scores—then identifies research where both approaches were examined. "Results and discussion" section presents a guideline proposal based on the results of the survey. "Conclusion" section concludes the paper and indicates future and related research directions.

Background

This section provides an overview of the most relevant topics related to text sentiment analysis. First the definition of Online Consumer Reviews (OCRs) is provided; then, sentiment analysis-related topics were identified and discussed.

Online consumer reviews (OCRs)

OCR is one of the most commonly used concepts to represent the traditional word-of-mouth review. An electronic word-of-mouth review, or OCR, is defined as "any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions

via the Internet" [24]. Word-of-mouth review, meaning personal opinions among people, has been recognized as a significant source of information to understand customers' interests, and sentiments concerning companies' products and services, such as movies, books, music albums, and enterprises such as hotels, and restaurants. Many consumers find word-of-mouth information to be useful and credible when making a decision about products or services because it is generated by independent pre-experienced consumers instead of biased company advertisements. With the rapid advancement of Internet technology, the electronic word-of-mouth technique has been adopted by different platforms such as Yelp, Amazon, and eBay to enable people to easily generate reviews, share them with other people, and exchange opinions. Electronic word-of-mouth information includes customer reviews, online comments, and score ratings, and it can be spread in real-time through online channels, such as e-commerce sites, online forums, the blogosphere, and social networking sites. Thus, electronic word-of-mouth information is recognized as not only a convenient way for consumers to share information, but also a source of new challenges and opportunities for business analysts to understand consumer interests and opinions.

Statistics also support reliance on Online Consumer Review (OCR) for the decision-making of consumers. Nearly 65% of consumers access consumer-written product reviews via the Internet [15]. Additionally, of those consumers who read reviews, 82% confirmed that reviews had directly influenced their buying decision, while 69% shared the reviews with others including: family, friends, and co-workers, so magnifying their impact. In addition, numerous surveys and consulting reports have suggested that, for a number of consumers and products (not applicable to all), consumer-generated reviews are valued more highly than reviews from 'experts' [21, 61]. Therefore, Online Consumer Reviews (OCRs) can impact the consumer decision-making process to a greater extent than traditional media [1].

Text mining and sentiment analysis

Data and text mining cover the broad scope of software tools and mathematical modelling approaches which are used to discover implicit, previously unknown patterns from data. In text mining, patterns are lifted from natural language text (unstructured data), while, in data mining, the patterns are lifted from structured databases. The text mining process starts with the text collection stage and then proceeds to the pre-processing stage in which the text is cleaned and formatted. The pre-processing stage involves critical tasks, such as tokenization, removal of stop words, and stemming. In the next stage, meaningful features are extracted to make inferences about the data. In the final stage, text mining approaches, such as categorization, topic modelling, or clustering, are applied to answer certain questions about the given data.

Large volumes of Online Consumer Reviews (OCRs) are available on many retailers' websites, and mining such data to understand consumers' opinions is called opinion mining. This term was first coined by Dave et al. [14], where opinion mining involves processing a set of reviews for a given product or service and extracting a list of attributes or aspects to categorize the consumer opinion into different classes such as (positive, negative, subjective, objective, etc.). Sentiment analysis is a subsection of opinion mining which focuses on the extraction of the consumer's emotions, opinions, and evaluations on services or products from online reviews they have posted. Sentiment analysis is an area of research that is very active, with a large volume of relevant research literature available [3]. Most of the research work is typically based on three common sentiment analysis tasks: subjectivity analysis, polarity detection, and sentiment strength detection. Subjectivity analysis aims to determine whether or not a given text is subjective, while polarity detection is utilized to assign an overall positive or negative sentiment orientation to subjective texts. Sentiment strength detection specifies the degree of polarity to which a text is either positive or negative.

Sentiment analysis is normally performed in two ways: a lexicon-based approach or a machine learning approach. A lexicon-based method uses a sentiment dictionary or a sentiment lexicon that is used to predict the overall sentiment of a text based on pre-defined word occurrence. Alternatively, a machine learning approach generates a classifying algorithm through learning with the set of linguistic features [28]. The trained classifier is then used for sentiment prediction [3].

In the lexicon-based approach, public lexicons, such as SenticNet [10], SentiWordNet [2], and OpinionFinder [63], have been frequently applied by many studies owing to the reliability of public sentiment dictionaries [28]. Lists of sentiment-bearing words and phrases available in opinion lexicon are used for lexicon-based techniques, such as the General Inquirer lexicon [54], WordNet Affect [55] SentiWordNet, the ANEW words [8], and the LIWC dictionary [43]. Beyond these standard resources and to automatically generate and score lexicons, researchers have created new methods. However, as indicated by Liu Y [34], while an opinion lexicon is required, it is insufficient for sentiment analysis. Thus, a combined approach is more appropriate as these approaches normally use additional information, e.g. semantic rules to handle emoticon lists, negation, booster word lists, and an already existing and substantial collection

of subjective logical statement patterns. According to Taboada et al. [56], "lexicon-based methods for sentiment analysis are robust, result in good cross-domain performance, and can be easily enhanced with multiple sources of knowledge."

LEX-Quality of Experience (QoE) parameters were utilized by [61] to analyse user reviews. The identification of frequent nouns in reviews was achieved through the utilization of speech tagging, and these were denoted as a prospective QoE element. Semantic lexicons, such as SentiWordNet, were used to group and aggregate similar nouns. For each group, the representative nouns were highlighted as QoE parameters. This work, therefore, exploited user reviews as inputs for quality element extraction from services through the selection of frequent nouns in drawing features and the end outcome.

Recently, machine learning algorithms have been used for most existing sentiment analysis techniques, such as Naive bayes (NB), support vector machines (SVM), neural network (NN), genetic algorithm (GA), and k-nearest neighbours (kNNs) to optimize, classify, and form predictions based on the data in text documents. Machine learning approaches have certain advantages, including the ability to identify the non-sentiment terms which express a sentimental judgement (e.g. "cheap" in the phrase "this camera is cheap"). An additional advantage of such approaches is the availability of a wide range of applicable learning algorithms. However, these methods present certain disadvantages, such as the need for a human-labelled corpus for the training phase. Additionally, while within the domain these trained machine learning methods perform very well, their performance can diminish significantly when applied to another domain. For example, in the cell phone domain, the words "cheap" and "smart" are used as expressions of positive opinions, while in the world of books domain, "well-researched" and "thriller" signify positive sentiments. Therefore, a cell phone domain-trained algorithm is unlikely to correctly classify book domain reviews. Moreover, as indicated by [58], some machine learning algorithms cannot "give a clear explanation as to why a sentence has been classified in a certain way, by reference to the predefined list of sentiment terms."

One can principally investigate sentiment analysis applications at three granularity levels: document level, sentence level, and aspect level. At the document level, the entire document is allocated an overall sentiment score. Sentence-level sentiment analysis concentrates on predicting the sentiment of stand-alone sentences. Subsequently, a score aggregation method is applied to generate an overall review score from combined sentence-level scores. However, in a document- or sentence-level analysis, it is not easy to obtain fine-grained opinions, though an aspect-level analysis can frequently overcome this problem. Aspect-level techniques carry out a finer-grained analysis with the intention of identifying sentiments on entities and/or their aspects [65].

Challenges in using sentiment analysis with OCRs

There are certain challenges and problems in implementing sentiment analysis, and some of them are as follows:

Short reviews: in a proposal by Cosma et al. [13] they state that in order to surmount the domain barrier in gathering views, there is need for an overall way of setting up language rules for the recognition of view-bearing words. Additionally, online reviews have unique text features that are short in length, use formless phrases, and involve substantial data. New challenges to conventional study topics in text analytics, i.e. text categorization, data mining, and emotional studies, are brought to the short reviews.

Colloquial language is another vital attribute of online text, specifically in online reviews. Consumers may use short forms or acronyms that rarely appear in traditional text when writing reviews, for example, phrases like "superb", "good 2go", hence making it extremely hard for one to identify the semantic meaning [5].

Mockery acknowledgement: the varied challenges require working through mockery or expressions that are unexpected. Riloff et al. [44] contributed to the improved review in mockery acknowledgement by developing an algorithm that naturally learned to group good and unpleasant phrases for tweets. The evaluation of two elements that are dissimilar amounts to analogy.

Domain dependency: the essential task of exploring the information generated by the customer lies in the wider concept of themes. Generally, the content generated is usually broad and needs to be packaged into categories. A classifier that is specified for a given domain space might thus not be effective in another domain which uses different words. The expression of sentiments is varied in different domains. This, notwithstanding the methods of sentiment categorization, can be synchronized to adequately work in a provided domain but, at the same time, may be limited to categorize sentiments in a varying domain. In light of this Bollegala et al. 2013 [7] proposed a cross-domain sentiment classifier that automatically extracts a sentiment thesaurus. Moreover, procedures or algorithms that are joined in a given area may not necessarily perform effectively in a space that is different from the initial one. The process of identifying specific area- and space-free systems was independently carried out. Jambhulkar and Nirkhi [25] carried out a cross-domain sentiment analysis survey study that focused on the following methods: sentiment-sensitive thesaurus, spectral feature alignment, and structural correspondence

learning. The findings of the study denoted that each of the used methods has its distinct way in (1) increasing the vector features, (2) evaluating the association between given words, as well as (3) the used classifier. According to Bisio et al. [4] there are two main features of notion characterization. These include the versatile nature of a provided structure and the subsequent ability to work in wider business spaces through utilization of relevant valence shifters, semantic systems, and a predictive model grounded on distance.

Methods

There are mainly three broad types to conduct a literature review including the systematic review, the semi-systematic review, and the integrative review [53]. For this research work, an integrative review of the literature was undertaken to critically analyse and examine the outcomes reported in related studies investigating different OCR solution models. An integrative review approach can be useful when the purpose of the review is not to cover all articles ever published on the topic but rather to combine different perspectives to effectively identify current problems and generate new knowledge about the topic [59]. In addition, identifying and analysing developed solution models using OCR is a broad topic, and a variety of disciplines such as business, marketing, and computing address various aspects of it. Therefore, we believe that using an integrative review would be a good choice for studying a broader topic that has been conceptualized differently and studied within diverse disciplines.

Databases such as Web of Science, Google Scholar, Scopus, and ScienceDirect have been accessed to search for existing research literature and documents relating to the topic. In addition, relevant research papers were accessed through backward citations of the articles included in the review. Relevant search terms: (text reviews, star ratings, score ratings, and text-ratings correlation) have been used to identify studies from 2002 to 2020. Then, to enhance the literature, we incorporated more keywords: (online reviews, product reviews, online recommendations, online word-of-mouth (e-WOM), online viral marketing, online consumer reviews, online communities, and virtual communities) to obtain articles from numerous management journals and relevant databases, including the Association for Computing Machinery Digital Library (ACM), IEEE Xplore, SCOPUS (Elsevier), and ScienceDirect (Elsevier). Papers' selection was structured in a two-stage process: first, excluding research studies based on reading the titles and the abstracts. In the second stage, research papers were filtered again from the initially selected list of papers, based on a complete reading. Around 66 papers with a minimum number of citations per paper 3 which were relevant, addressing the research questions, and contributing to the basic purpose of the review have been included in the review.

An ontology to conceptualize knowledge in the domain of text reviews has been proposed. Protégé [38] was used to build the ontology that includes the main concepts in the domain of review and review analysis. The ontology also determines the relationships between the concepts. Figure 2 shows the proposed review ontology which consists of 31 classes to conceptualize and classify the concepts of review domain based on the analysis of this review.

Related work

Reviewed papers are broadly first categorized into two themes: (1) research studies that built their models based on either rating scores or review texts, assuming that they were correlated implicitly with each other. In this part, we included research studies that have not considered any validation metrics to compute the relationship degree between available rating scores and other labelling techniques. They only use one source of labelling, only

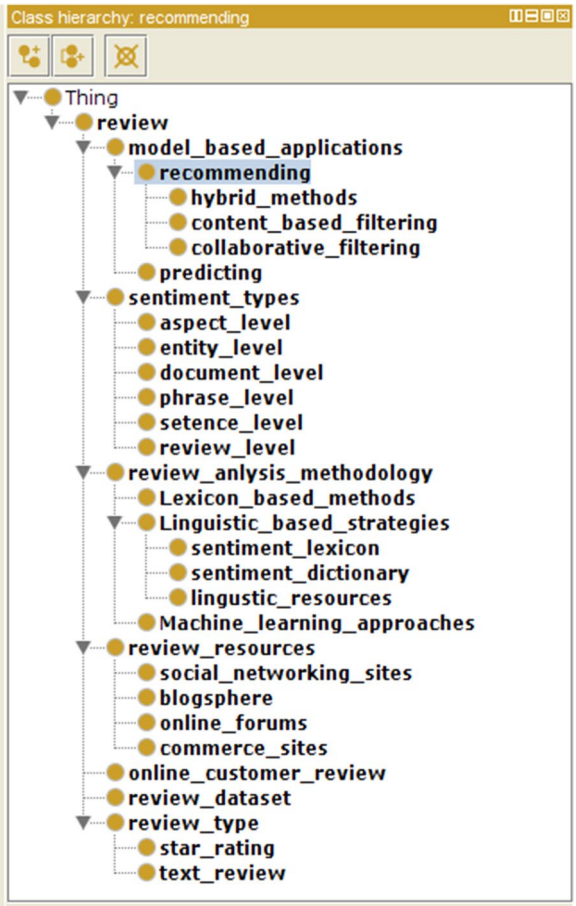

Fig. 2 Review ontology

numeric ratings, or review texts' sentiments, to build their OCR solution models. (2) Research work that examined the correlation between the text reviews' sentiments and the associated rating scores before applying the proposed solution model. Hence, numeric rating scores have been validated using other labelling techniques such as review texts annotation using either experts or sentiments lexicons.

Then, secondly, for each categorized theme, all relevant literature was classified into categories based on the following criteria: research studies use OCR on survey analysis, research studies use OCR for prediction/ classification models, and research studies use OCR for recommendation models. Summary tables for the reviewed research studies are presented at the end of each subsection. We examined the reviewed papers based on the following considerations: used model, domain, manually labelled (reviews were labelled by experts), and automatic labelling (reviews were labelled by sentiment lexicons). For research studies that examined the correlation between the text reviews' sentiments and the rating scores, we checked, as well, the degree of consistency between text reviews and rating scores. Figure 3 shows the adopted framework for conducting the review. All relevant literature used in this paper are provided in Appendix.

1. Numeric rating labels have not been validated against other labelling techniques:

In surveying the literature, we found that many research studies have not considered any validation metrics to compute the relationship degree between available rating scores and other labelling techniques, whether they were survey papers or sentiment classification using machine learning algorithms or other different techniques. In the context of online reviews survey analysis, Shoham et al. [51] examined the effect of irrelevant reviews and its associated positive or negative rating scores on customers' product evaluations and future decisions. The survey analysis study was done using 7913 reviews of approximately 100 products in five different classes. The survey results revealed that the presence of irrelevant reviews with negative rating scores alongside positive reviews leads to greater product preferences, as consumers feel confident that the information they have about the product is more complete. This finding suggests that sellers or service providers should not be discouraged by negative or totally unhelpful, irrelevant reviews, or attempt to block customers from seeing them.

For research studies that apply machine learning (ML) algorithms for sentiment classification, Kim, Kwon, and Jeong [28] detected the availability of machine learning models only using linguistic features and identified the influence of the size of the linguistic feature set on the classification accuracy. They conducted sentiment analysis study focusing on Korean electronic word-of-mouth (eWOM) in the film market and selected 10,000 movie reviews, which were rated with negative and

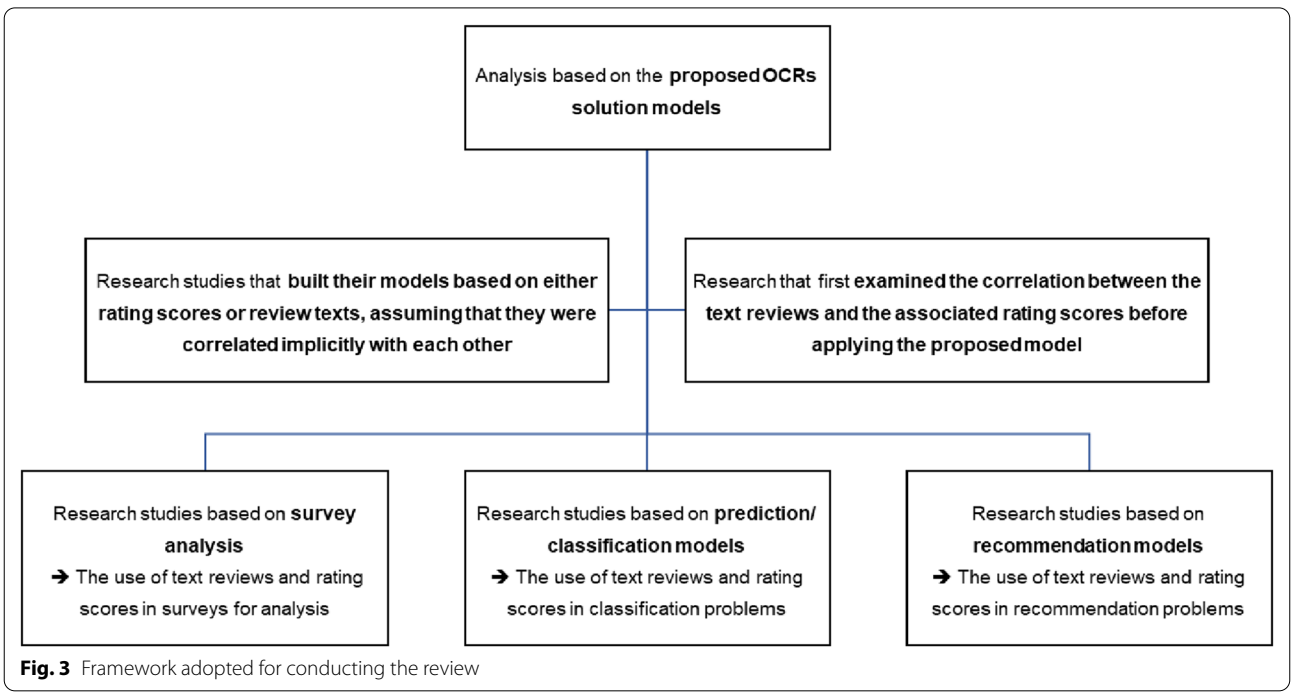

Fig. 3 Framework adopted for conducting the review

positive popularity on the movie portal sites, and parsed words through natural language processing (NLP). Four machine learning methods: Naïve bayes (NB), decision tree (DT), neural network (NN), and support vector machine (SVM) were demonstrated with the linguistic features, and their performances were compared by accuracy and the harmonic mean between precision and recall (F1 score). In addition, Kim et al. tested five different feature set sizes groups to see whether the feature set size influenced the performance of classification. As a result, neural network (NN) and support vector machine (SVM) classification showed acceptable performance under every condition. Through the experiments, Kim et al. showed how machine learning algorithms are applied as sentiment classifiers for movie electronic word-of-mouth (eWOM) analytics, and a performance gap might have occurred with this method as a result of the feature set size. Another study done by Rui, Liu, and Whinston [46] employed support vector machine (SVM) and Naïve bayes (NB) classifiers to assess word-of-mouth (eWOM) impact of consumer's willingness. They categorized 4,166,623 movie tweets into four mutually exclusive categories: intention, positive, negative, and neutral. Researchers trained NB for intention and support vector machine (SVM) for sentiment, and validated word-of-mouth (eWOM) impact of tweets with precision and recall as performance measures.

Hamouda et al. [22] proposed also a machine learning senti-word lexicon based on training support vector machine (SVM) algorithm using Amazon corpus containing reviews from various domains. They converted Amazon reviews dataset into binary classes (positive and negative) by assigning reviews with a 1 or 2 rating scores as negative reviews and the reviews with a 4 or 5 rating scores as positive reviews, while ignoring reviews with a 3-star rating scores. They have provided an upgrade for creating a lexicon by using 'Strong Reviews' for the dataset and 'root' of tokens as the linguistic feature used by support vector machine (SVM). An additional improvement in the accuracy of reviews classification comes from using 'Term Score Summation' for sentiment computation. Pang et al. [40] applied three machine learning methods Naïve bayes (NB), support vector machine (SVM), and maximum entropy (MaxEnt) to determine whether a movie review was positive or negative. A corpus of 752 negative and 1301 positive reviews, with a total of 144 reviewers from Internet Movie Database (IMDb) archive, has been used for the study. They examined several features conditions, such as n-gram, parts of speech (POS), and position of the word. They achieved a best performance using 16,165 unigrams features and support vector machine (SVM) method by accounting only for feature presence. Interestingly, their experiment also showed that using top frequent 2633 unigrams, accuracy was very similar to the best performance noted above. This means that a small-sized feature set can be considered as an efficient method for sentiment analysis of big data [28].

Prediction of review numeric rating scores is one of the main tasks of sentiment analysis, and it stretches the binary sentiment classification and focuses more on predicting the numeric rating (e.g. 10 stars) for a given review. Pang and Lee 2005 [39] looked into prediction of a review rating as a classification regression challenge; therefore, they created a rating predictor with machine learning method under a supervised metric labelling framework. They proposed a meta-algorithm using metric labelling to ensure that similar items receive similar labels. The results showed that the proposed model outperformed both multi-class and regression versions of support vector machine (SVM).

Through taking into account user information, Tang et al. [57] proposed a neural network method for review rating prediction. In their paper, they targeted a finer-grained document-level problem and conducted experiments on two benchmark datasets, Yelp13 for restaurant reviews and RT05 for movie reviews. They used two main models: the user-word vector model which modifies the original word vectors with user information, and the document vector model which takes the modified word vectors as input and produces review representation that are used as the feature for predicting review rating. The proposed method marginally outperformed text-based neural network algorithms convolutional neural network (CNN), for the following datasets: Yelp and RT, as they captured user-level and text-level semantics simultaneously.

Basiri et al. [3] detected the polarity of reviews by adopting a machine learning technique, and then, they considered sentence scores as proof for overall review ratings. In order to predict review scores, they first discover the individual sentences' scores within a review and then group them into five-star review scales. To detect emotions at the sentence level, they used SentiStrength, an available library for lexicon-based sentiment strength detection. Experiments were carried out on CitySearch for restaurants reviews and TripAdvisor for hotels reviews. The results showed that the proposed model outperforms the existing aggregation methods with regard to accuracy and mean absolute error (MAE). However, the proposed model does not perform well compared to some machine learning algorithms such as AdaBoost, Bayesian Networks, Decision Tree (DT), K-Star, Naive bayes (NB), and support vector machine (SVM) in terms of accuracy. The main advantage of the proposed model is that it outperforms other machine

learning algorithms in terms of speed and memory requirements. Table 1 summarizes the reviewed papers that did not consider any validation procedure to examine the relationship between texts reviews and numeric ratings.

2. Numeric rating labels have been validated against other labelling techniques:

In this subsection, a sample of research studies that have examined the relation between numeric ratings and text reviews is discussed and reviewed. Starting with Zhu et al. [68] who examined the link between guests' text reviews and score ratings in the tourism and hospitality domain, based on the text reviews of 4602 Airbnb accommodation listed in San Francisco, USA, the main finding was there is a strong relationship between the positive (negative) sentiment and the high (low) score ratings. People tend to give higher score rating for positive reviews and low score rating for negative reviews. They applied the Tobit model and results indicate that there is a higher rating score can be expected if the guest's comment was more positive and the opposite for the negative reviews.

Li [29] examined review reliability by using sentiment analysis which was based on reviews left by travellers with Skytrax and connected Twitter messages. This study examined the extent to which sentiments within reviews about air experiences with Skytrax correlated to the Star-Airline Ratings (1–5), and how travellers' feelings on air travel experiences differed on Skytrax to those left on Twitter. Results showed that the Airlines Rating programme (1–5 stars) actually had a low level of reliability based on what airlines knew what had been posted on Twitter. Two tests revealed that there is a nominal positive correlation between sentiments within reviews from Skytrax compared to Star-Airlines ratings (20.7%). In addition, the Airlines Rating programme clearly reveals a fragile external validity. Although text sentiments from Skytrax and Twitter had a positive correlation, the level had statistical significance. In total, 4033 Skytrax reviews were used for the analysis in addition to 10,522 tweets, and related comments for 177 airlines were gathered by individually searching under each airline's unique name.

Geetha et al. [18] investigated the relationship between customer online review sentiments and guests' hotels ratings. They examined if the customer sentiment polarity had a positive effect on their ratings. Results that have shown consistency between customer reviews and hotel ratings are not entirely consistent across budget and premium hotel categories. It explains the sense of 44% of the variance in the customer rating for budget category and 21% of the variance in customer rating for premium category. They found that there is a linear relationship between customer rating and customer sentiment polarity.

Tsang and Prendergast [60] investigates how the inconsistency of positive or negative reviews between text reviews and ratings affects the consumers and shows that there is a link between text reviews and ratings. It was found that text positively or negatively significantly influenced consumers' reactions of reviews. Sellers can benefit by incorporating both text reviews and ratings to enhance the prediction accuracy of the products' sales. In

Table 1 Summary of the reviewed papers in which no validation procedure has been applied to examine the relationship between numeric ratings and texts reviews

	Reviewed Paper	Used Model	Domain	(Manually labelling) Reviews were labelled by Experts?	(Automatic labelling) Reviews were labelled by Sentiment Lexicons?
Survey Papers	Shoham et al. [51]	Survey Study	–	–	–
Sentiment Analysis using Machine Learning	Kim et al. [28]	Binary classification	Movies	No	No
	Rui et al. [46]	Multiclass classification	Movies	No	Yes
	Hamouda et al. [22]	Binary classification	Amazon from different domains (books, cameras, mp3s, etc.)	No	No
	Pang and Lee [39]	Multiclass classification Regression	Movies	Yes	No
	Pang et al. [40]	Binary classification	Movies	No	No
Sentiment Analysis using Different Techniques	Tang et al. [57]	Neural Network + user-word composition vector model (UWCVM)	Restaurants and movies	No	No
	Basiri et al. [3]	Lexicon-based model	Restaurants and hotels	Partially	Yes

the survey results, they collected 30 responses for each of the 24 releases which formed a sample size of 720. They collected the data in three high-traffic areas in Hong Kong and assigned the reviews to participants randomly. After reading them, their understanding was measured. Then, they conducted the manipulation checks and collected their demographic information.

Ganu et al. [16] compared the users' star rating with text reviews using Pearson correlation coefficient which ranges from −1 to 1. Using a corpus contains 5531 restaurants, with associated a set of 52,264 reviews. Reviews contain structured metadata (star rating, date) along with text. The experiment showed there was a positive correlation between positive reviews and star ratings and a negative correlation between negative reviews and star ratings. These results motivated the authors to include text reviews in the context of recommender systems. Research hypothesis is that the review text is a better indicator of the review than the coarse star rating. They test this hypothesis in the recommendation system scenario and explore whether text-derived ratings are better predictors than numerical star ratings given a user's restaurant preferences.

In addition to the application on prediction models, considering both score ratings and text review plays a vital role in recommendation systems. In order to recommend products to users we must finally predict how the user responds to a new product. To do this, we must disclose the implicit tastes of each user as well as the characteristics of each product. For example, in order to predict whether a user will enjoy Harry Potter, it helps to determine that the book is about wizards, as well as the level of the user's interest in wizardry. User feedback is required to discover these inherent dimensions of the product and the user. This feedback often comes in the form of a numeric rating accompanied by the review text. However, traditional methods often ignore review text, making it difficult to fully interpret user and product dimensions, as they discard the same text that justifies a user's rating.

Yu et al. [67] proposed a transformation that links the users' or items' average rating with sentiment probability to better rating prediction. They transform the average rating of items to the sentiment distribution in the text reviews and map the average rating score into a probability space of the sentiment distribution. Using a real dataset from Amazon, they found that mean squared error (MSE) using their model achieved the smallest one, i.e. 1.361, and thus performed the best among all considered models. Ling et al. [31] proposed a generative model that combines a topic model with a rating model. Experiments show that the proposed model leads to significant improvement compared with strong baseline methods, especially for sparse datasets where rating-only methods cannot make accurate predictions (cold-star setting).

The researchers McAuley and Leskovec [35] indicated that most of the research work in the domain of reviews and rating were studied disjointedly. Therefore, the authors proposed a methodology for predicting reviews accurately and for genre automated discovery by combining both text reviews and rating. In addition, the authors pointed out that the research area for studying reviews includes understanding the rating process and predicting rating. In their paper, the authors found out that predicting review accuracy can be increased by combining text and review.

Tables 2 and 3 summarize the reviewed papers which use a validation tool to examine the relationship between texts reviews and numeric ratings for both prediction and recommendation models, respectively.

Results and discussion

After surveying the literature, in most of the reviewed papers, sentiment labels are obtained from the review text or the associated rating scores. We argue that there are differences between the sentiments of the reviewed text and the associated numeric ratings and ought to be considered. This issue has been largely ignored and only some studies such as [18, 67, 68] have partially taken it in their considerations while building their solutions. In addition, reviewed research papers that assessed the relationship between text reviews and the associated rating scores have revealed low to average correlations. This analysis result suggests that building solution models based on only the texts' sentiments or the numeric rating scores should be used with caution in practice.

It should be noted that this research paper identifies the problem of inconsistency between text review and numeric scoring and how this might question its usefulness as labels for building OCRs solution models. Hence, a non-tested correlation or even weak correlation suggests or implies for inaccurate labels which affect the model outputs. Otherwise stated, when developed models learn from inaccurate labels, they output inaccurate predictions and recommendations. Other research work has discussed discrepancies sources and indicated that review texts do not correlate well with the review outcomes may be results of random errors or the subjective process involved in presenting the review. The source of discrepancies has been examined by Geierhos et al. [19] and Jang and Park [26]. In this regard, Geierhos et al. [19] pointed out that one of the reasons for the inconsistency is individual random errors, while Jang and Park [26] attributes this position to two possible sources of uncertainty: reference uncertainty (reviewers are affected by previous reviews) and reference heterogeneity (reviewers

Table 2 Summary of the reviewed papers which examine the relationship between numeric ratings and texts reviews in prediction models

Paper	Used Model	Domain	(Manually labelling) Linking between Experts' ratings and available rating scores were checked?	(Automatic labelling) Linking between lexicon-based review ratings and available rating scores were checked?	Degree of consistency between text reviews and rating scores
Zhu et al. [68]	Regression Model	Tourism and hospitality (Airbnb)	No	Yes Tobit Models	Coefficients = 0.3072 for positive and −4.2846 for negative Degree: Weak for positive
Li [29]	Statistical Model	Airlines	No	Yes "Semantria" Kendall's tau and the Spearman's rho Correlation	Kendall's tau = 0.207 and the Spearman's rho = 0.268 Degree: Weak
Geetha et al. [18]	Naive Bayes Classifier	Hotels	No	Yes Linear Regression Model	A linear relationship between customer sentiment and rating R square has a value of 0.21. So, 21% of the variation in the customer ratings is explained by customer sentiment polarity Degree: Weak
Tsang And Prendergas, 2009 [60]	Statistical Study	Movies	No	Yes Analysis of covariance (ANCOVAs)	A significant interaction between text and rating valences on trustworthiness
Ganu et al. [16]	Support Vector Machine classifiers (sentiment classification) Regression (numeric scores)	Restaurants	Yes	Yes Private lexicon done by researchers Percentage Analysis	56% of the reviews were annotated as positive and 18% as negative. The associated numeric ratings provided by users pointed that 73% of reviews having positive rating Degree: Average

Table 3 Summary of the reviewed papers which incorporate both numeric ratings and texts reviews in recommendation models

Paper	Datasets	Incorporating both users' numeric ratings and users' text reviews?
Yu et al. [67]	Amazon Review dataset (Arts, Jewelery, Watches, Cell Phones and Accessories, etc.)	Yes They mapped between aspects sentiments in review texts and rating scores to better rating predication
Ling et al. [31]	Amazon Review dataset (Arts, Jewelery, Watches, Cell Phones and Accessories, etc.)	Yes They applied topic modelling techniques on the review text and aligned the topics with rating dimensions to improve prediction accuracy. They were able to improve the accuracy over existing strong baseline methods, that use only rating for recommendations especially under the cold start problem when the data is extremely sparse
McAuley and Leskovec [35]	Amazon Review dataset (e.g. Books, Movies) + pub data from ratebeer.com + restaurant data from citysearch.com, + Yelp dataset	Yes They proposed a model that works by aligning hidden factors in product ratings with hidden topics in product reviews. The proposed model allows to accurately fit user and product parameters with only a few reviews, which existing models cannot achieve using only a few ratings

have different backgrounds and experiences). Mellinas et al. [37] and Sharma et al. [49] in their analysis concluded that customers tend to punish dissatisfaction more harshly than satisfaction.

Here in this work, we propose some of the guidelines that can help reduce the effect of discrepancy between text reviews and ratings. The following subsections introduce guidelines we derived using the results of our study.

Proposed guideline to incorporate text reviews and numeric ratings

To overcome the available Text-Rating Review Discrepancy (TRRD) shortcoming, we suggest measuring text reviews and ratings correlation and consider their agreement and disagreement level into account. For example, to build a model, which predicts the review rate for a given text, we could select the training data instances (text reviews) that have an agreement between the experts' annotations, numeric ratings, and sentiment lexicon results. To build prediction models, we propose the following steps shown in Fig. 4. It starts with annotating data from different sources such as experts, numeric ratings, and sentiment lexicons. Then, a measure of agreement should be used to reflect the amount of agreement for the annotated reviews using different labelling sources. Finally, the review instances that have a strong correlation between its annotation values using different methods would be selected for building the model that predicts the review rate for a given text.

In the case of the recommender models, we also propose to incorporate both text reviews and numeric rates in order to understand users' preferences and interests and therefore deliver better customer service. Figure 5 illustrates the proposed model for building recommendation models. It should be noted that the illustrated steps in Figs. 4 and 5 are the basic steps of the development of any prediction/recommendation model. The only added phase was to ensure the agreement between text reviews and associated numeric ratings.

Validation measures

In this study, we propose to examine the correlation between available rating scores, rating scores provided by annotators, and scores calculated using lexicon-based models. The proposed parallel set of measures to compute the used labels' validity are listed in Table 4. The provided evaluation framework encompasses a set of quantitative measures that provides estimated results of the instances' labels' validity.

Conclusion

The use of Online Consumer Reviews (OCRs) has attracted researchers across multiple disciplines such as business, marketing, and computing. In most proposed solutions for analysing the online customers reviews, rating scores, and review texts were the primary components that have been employed to produce high-quality Online Consumer Reviews (OCRs) solution models. However, most of the reviewed models consider either rating scores or the review texts to build their models assuming that both were correlated with each other. This paper introduces the concept Text-Rating Review Discrepancy (TRRD) which is defined by the inconsistency between text reviews and score ratings for a product

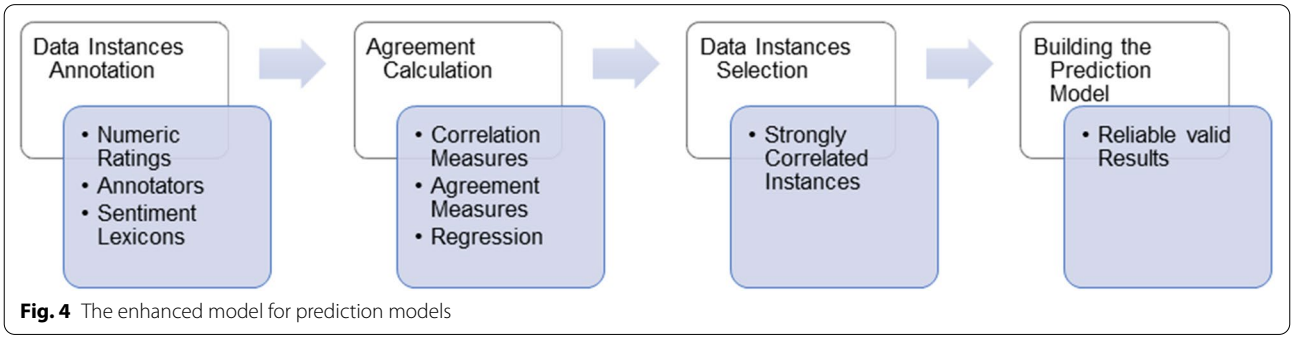

Fig. 4 The enhanced model for prediction models

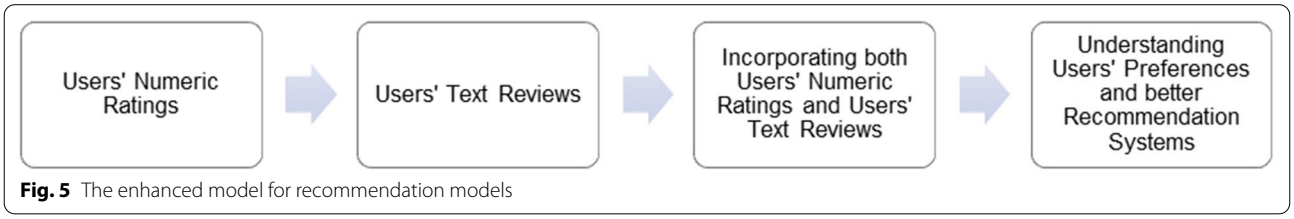

Fig. 5 The enhanced model for recommendation models

Table 4 Quantitative measures to compute the relationship degree between available rating scores and other labelling

Measure	Purpose
Similarity and Consistency	To determine whether available rating scores correlate with other labelling techniques (lexicon and expert annotators labelling) Example: cosine similarity, Pearson's correlation coefficient, Spearman's rho, etc.
Agreement	To determine whether available rating scores agrees with other labelling techniques (lexicon and expert annotators labelling) Example: per cent agreement, alpha agreement, etc.
Linear Regression	To determine whether a linear relationship exists between available rating scores and other labelling techniques (lexicon and expert annotators labelling)

or service posted review. The main contribution of this paper includes showing the necessity for using both text reviews and score ratings to ensure having reliable survey results and building valid models. We therefore reviewed the literature to identify if there are any discrepancies between text reviews and numeric ratings. In surveying the literature, we found that research studies that assessed the relation between text reviews and the associated rating scores have revealed low to average correlations. This finding suggests that building solution models based on only the texts' sentiments or the numeric rating scores should be used with caution in practice. Alternatively stated, the presented exploratory analysis shows that customers might express text sentiments which are different from the associated numeric rating scores. Therefore, we propose to take full advantage of the abundant information of the text reviews sentiments and examine its relationship degree to the combined rating scores. Then, employ the most correlated data instances in order to build a more accurate model. Our research suggests that sentiment of a review combined with a correct numeric rating would be an indicator for the validity and correctness of the required OCR solution model. Also, this study encourages researchers to look beyond the numeric ratings into the text sentiments as written texts can express more information and emotions which quantitative ratings cannot capture. To end with, future research should attempt to ensure the correctness and quality of both the text review and associated numeric ratings. In addition, it should also pay more attention to the causes of discrepancies and inconsistencies between text reviews and ratings in order to mitigate and reduce its negative effects on developed OCRs solution models.

Appendix

The following table summarizes both: (1) research studies that built their models based on either rating scores or review texts, assuming that they were correlated implicitly with each other. (2) research studies that examined the correlation between the text reviews' sentiments and the associated rating scores before applying the proposed solution model.

Numeric Rating Labels Have Not Been Validated Against Other Labelling Techniques	
Paper Author(s)	**Paper Title**
Basiri et al. [3]	Sentiment prediction based on Dempster-Shafer theory of evidence
Hamouda et al. [22]	Building machine learning based senti-word lexicon for sentiment análisis
Kim et al. [28]	[Comparing machine learning classifiers for movie WOM opinion mining
Pang and Lee [39]	Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales
Pang et al. [40]	Thumbs up? Sentiment classification using machine learning techniques
Rui et al. [46]	Whose and what chatter matters? The effect of tweets on movie sales
Shoham et al. [51]	Positively useless: Irrelevant negative information enhances positive impressions
Tang et al. [57]	User modeling with neural network for review rating prediction

Numeric Rating Labels Have Been Validated Against Other Labelling Techniques	
Paper Author(s)	**Paper Title**
Ganu et al. [16]	Beyond the stars: Improving rating predictions using review text content
Geetha et al. [18]	Relationship between customer sentiment and online customer ratings for hotels: An empirical analysis
Li [29]	Application of sentiment analysis: Assessing the reliability and validity of the global airlines rating program
Ling et al. [31]	Ratings meet reviews, a combined approach to recommend
McAuley and Leskovec [35]	Hidden factors and hidden topics: Understanding rating dimensions with review text
Tsang and Prendergas [60]	Is a "star" worth a thousand words?: The interplay between product-review texts and rating valences
Yu et al. [67]	Rating prediction using review texts with underlying sentiments

Numeric Rating Labels Have Been Validated Against Other Labelling Techniques

Paper Author(s)	Paper Title
Zhu et al. [68]	Sentiment and guest satisfaction with peer-to-peer accommodation: When are online ratings more trustworthy?

Abbreviations

CNN: Convolutional neural network; DT: Decision tree; GA: Genetic algorithm; kNNs: K-nearest neighbours; MAE: Mean absolute error; MaxEnt: Maximum entropy; MSE: Mean squared error; ML: Machine learning; NB: Naive bayes; NLP: Natural language processing; NN: Neural network; OCR: Online consumer review; QoE: Quality of experience; SVM: Support vector machines; TRRD: Text-rating review discrepancy; WOM: Word-of-mouth.

Acknowledgements

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant Number (DF-110-165-1441). The authors, therefore, gratefully acknowledge DSR technical and financial support.

Authors' contributions

AM, RO, and HH contributed to the design and execution of the research review, to the analysis of the results, and to the writing of the manuscript. All authors have read and approved the manuscript.

Funding

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant Number (DF-110-165-1441). The authors, therefore, gratefully acknowledge DSR technical and financial support.

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no Competing interests.

Received: 27 September 2021 Accepted: 25 January 2022
Published: 22 February 2022

References

1. Anand O, Srivastava PR, Rakshit A (2017) Assessment, implication, and analysis of online consumer reviews: A literature review. Pacific Asia Journal of the Association for Information Systems, vol 9, no 2
2. Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Language resources and evaluation conference, pp 2200–2204.
3. Basiri ME, Naghsh-Nilchi AR, Ghasem-Aghaee N (2014) Sentiment prediction based on Dempster-Shafer theory of evidence, Mathematical Problems in Engineering
4. Bisio F, Gastaldo P, Peretti C, Zunino R, Cambria E (2013) Data-intensive review mining for sentiment classification across heterogeneous domains. In: IEEE/ACM International conference on advances in social networks analysis and mining (ASONAMI), pp 1061–1067. https://doi.org/10.1145/2492517.2500280
5. Blenn N, Charalampidou K, Doerr C (2012) Context-sensitive sentiment classification of short colloquial text. In International Conference on Research in Networking, Berlin, Heidelberg: Springer, pp 97–108. https://doi.org/10.1007/978-3-642-30045-5_8
6. Bloem C (2017, July). 84 percent of people trust online reviews as much as friends. Here's how to manage what they see. Inc. [Online]. Available: https://www.inc.com/craig-bloem/84-percent-of-people-trust-online-reviews-as-much-.html
7. Bollegala D, Weir D, Carroll J (2013) Cross-domain sentiment classification using a sentiment sensitive thesaurus. IEEE Trans Knowl Data Eng 25(8):1719–1731. https://doi.org/10.1109/TKDE.2012.103
8. Bradley MM, Lang PJ (1999) Affective norms for English words (ANEW): Instruction manual and affective ratings, the center for research in psychophysiology, University of Florida, Technical report C-1, vol 30, no 1, pp 25–36.
9. BrightLocal. (2019, December). Local consumer review survey: Online reviews, statistics & trends. [Online]. Available: https://www.brightlocal.com/research/local-consumer-review-survey/
10. Cambria E, Olsher D, Rajagopal D (2014) SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In: AAAI conference on artificial intelligence, pp 1515–1521.
11. Chang YC, Ku CH, Chen CH (2019) Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. Int J Inf Manage 48:263–279. https://doi.org/10.1016/j.ijinfomgt.2017.11.001
12. Nielsen. (2012, November). Consumer trust in online, social and mobile advertising grows. [Online]. Available: http://www.nielsen.com/us/en/insights/news/2012/consumer-trust-in-online-social-and-mobile-advertising-grows
13. Cosma AC, Itu VV, Suciu DA, Dinsoreanu M, Potolea R (2014) Overcoming the domain barrier in opinion extraction. In: 2014 IEEE international conference on intelligent computer communication and processing (ICCP), pp 289–296
14. Dave K, Lawrence S, Pennock DM (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: ACM 12th international conference on the world wide web, pp 519–528.
15. Consumer Goods. (2007, April) Deloitte study shows inflection point for consumer products industry. [Online]. Available: https://consumergoods.com/deloitte-study-shows-inflection-point-consumer-products-industry
16. Ganu G, Elhadad N, Marian A (2009) Beyond the stars: Improving rating predictions using review text content. In: 12th international workshop on the web and databases, pp 1–6.
17. Ganu G, Kakodkar Y, Marian A (2013) Improving the quality of predictions using textual information in online user reviews. Inf Syst 38(1):1–15
18. Geetha M, Singha P, Sinha S (2017) Relationship between customer sentiment and online customer ratings for hotels: an empirical analysis. Tour Manage 61:43–54
19. Geierhos M, Bäumer S, Schulze, Stuß V (2015) I grade what I get but write what I think, Inconsistency Analysis in Patients' Reviews" (2015). In: European Conference on Information Systems
20. Goldberg AB, Zhu X (2006) Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In: The first workshop on graph-based methods for natural language processing, TextGraphs-1, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 45–52.
21. Gu X, Kim S (2015) What part of your apps are loved by users? (T). In: 2015 30th IEEE/ACM international conference on automated software engineering (ASE), Lincoln, NE, USA, pp 760–770. https://doi.org/10.1109/ASE.2015.57
22. Hamouda A, Marei M, Rohaim M (2011) Building machine learning based senti-word lexicon for sentiment analysis. J Adv Inf Technol. https://doi.org/10.4304/jait.2.4.199-203
23. Hearn A (2010) Structuring feeling: Web 2.0, online ranking and rating, and the digital 'reputation' economy. Ephemera Theory Polit Org 10:421–438
24. Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? J Interact Mark 18(1):38–52
25. Jambhulkar P, Nirkhi S (2014) A survey paper on cross-domain sentiment analysis. Int J Adv Res Comput Commun Eng 3(1):5241–5245
26. Jang W, Kim J, Park Y (2014) Why the online customer reviews are inconsistent? Textual review vs. scoring review, Digital Enterprise Design & Management. Advances in Intelligent Systems and Computing, vol 261, pp 151–151. https://doi.org/10.1007/978-3-319-04313-5_20

27. Khalid H, Shihab E, Nagappan M, Hassan AE (2015) What do mobile app users complain about? IEEE Softw 32(3):70–77. https://doi.org/10.1109/MS.2014.50
28. Kim Y, Jeong SR (2015) Comparing machine learning classifiers for movie WOM opinion mining. KSII Trans Internet Inf Syst 9(8):3169–3181
29. Li G (2017) Application of sentiment analysis: Assessing the reliability and validity of the global airlines rating program, B.S. thesis, University of Twente.
30. Li YM, Chen HM, Liou JH, Lin LF (2014) Creating social intelligence for product portfolio design. Decis Support Syst 66:123–134. https://doi.org/10.1016/j.dss.2014.06.013
31. Ling G, Lyu MR, King I (2014) Ratings meet reviews, a combined approach to recommend. In: Proceedings of the 8th ACM conference on recommender systems (RecSys), New York, NY, USA, pp 105–112. https://doi.org/10.1145/2645710.2645728
32. Lipizzi C, Iandoli L, Marquez JER (2015) Extracting and evaluating conversational patterns in social media: a socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. Int J Inf Manage 35(4):490–503
33. Lipsman A (2009, November). Online consumer-generated reviews have significant impact on offline purchase behavior. Comscore. [Online]. Available: https://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior
34. Liu Y (2006) Word of mouth for movies: its dynamics and impact on box office revenue. J Mark 70(3):74–89
35. McAuley J, Leskovec J (2013) Hidden factors and hidden topics: Understanding rating dimensions with review text. In: Proceedings of the 7th ACM conference on recommender systems, pp 165–172.
36. Mcilroy S, Shang W, Ali N, Hassan AE (2017) User reviews of top mobile apps in Apple and Google App stores. Commun ACM 60(11):62–67. https://doi.org/10.1145/3141771
37. Mellinas JP, Nicolau JL, Park S (2019) Inconsistent behavior in online consumer reviews: the effects of hotel attribute ratings on location. Tour Manage 71:421–427
38. Musen MA (2014) The Protégé project: a look back and a look forward. AI Matters 1(4):4–12. https://doi.org/10.1145/2757001.2757003
39. Pang B, Lee L (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: 43rd annual meeting of the Association for Computational Linguistics (ACL), Ann Arbor, MI, USA, pp 115–124. https://doi.org/10.3115/1219840.1219855
40. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: ACL-02 conference on empirical methods in natural language processing (volume 10), pp 79–86.
41. Panichella S, Di Sorbo A, Guzman E, Visaggio CA, Canfora G, Gall HC (2015) How can I improve my app? Classifying user reviews for software maintenance and evolution. In: 2015 IEEE international conference on software maintenance and evolution (ICSME), IEEE Computer Society, Washington, DC, USA, pp 281–290. https://doi.org/10.1109/ICSM.2015.7332474
42. Park DH, Liu M, Zhai C, Wang H (2015) Leveraging user reviews to improve accuracy for mobile app retrieval. In: 38th International ACM SIGIR conference on research and development in information retrieval (SIGIR), ACM, New York, NY, USA, pp 533–542. https://doi.org/10.1145/2766462.2767759
43. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: LIWC 2001, Mahway: Lawrence Erlbaum Associates, vol 71, no. 2001
44. Piller C (1999, December). Everyone is a critic in cyberspace. Los Angeles Times. [Online]. Available: http://articles.latimes.com/1999/dec/03/news/mn-40120
45. Riloff E, Qadir A, Surve P, Silva LD, Gilbert N, Huang R (2013) Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 704–714.
46. Rui H, Liu Y, Whinston A (2013) Whose and what chatter matters? The effect of tweets on movie sales. Decis Support Syst 55(4):863–870. https://doi.org/10.1016/j.dss.2012.12.022
47. Salakhutdinov R, Andriy M (2008) Bayesian probabilistic matrix factorization using markov chain monte carlo. In: 33rd international conference on machine learning, pp 880–887.
48. Shan G, Zhang D, Zhou L, Suo L, Lim, J, Shi C (2018) Inconsistency investigation between online review content and ratings. In: Americas Conference on Information Systems (AMCIS).
49. Sharma A, Park S, Nicolau J (2020) Testing loss aversion and diminishing sensitivity in review sentiment, Tourism Management, vol 77, p 104020
50. Shelke N, Deshpande S, Thakare V (2017) Domain independent approach for aspect-oriented sentiment analysis for product reviews. In: 5th international conference on frontiers in intelligent computing: Theory and applications, Springer, pp 651–659.
51. Shoham M, Moldovan S, Steinhart Y (2017) Positively useless: Irrelevant negative information enhances positive impressions. J Consum Psychol 27(2):147–159. https://doi.org/10.1016/j.jcps.2016.08.001
52. . Singh K, Piryani R, Uddin A, Waila P (2013) Sentiment analysis of movie reviews: a new feature-based heuristic for aspect-level sentiment classification. In: 2013 international multi-conference on automation, computing, communication, control, and compressed sensing (IMac4s), IEEE, pp 712–717
53. Snyder H (2019) Literature review as a research methodology: an overview and guidelines. J Bus Res 104:333–339
54. Stone PJ, Dunphy DC, Smith MS (1966) The general inquirer: a computer approach to content analysis. MIT Press, Oxford, England
55. Strapparava C, Valitutti A (2004) WordNet affect: an affective extension of WordNet. In: 4th International conference on language resources and evaluation (LREC)
56. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267–307
57. Tang D, Qin B, Liu T, Yang Y (2015) User modeling with neural network for review rating prediction. In: 24th international conference on artificial intelligence (IJCAI), AAAI Press, pp 1340–1346.
58. Thelwall M, Buckley L, Paltoglou G, Skowron M, Garcia D, Gobron S, Ahn J, Kappas A, Küster D, Holyst JA (2013) Damping sentiment analysis in online communication: Discussions, monologs and dialogs. In: International conference on intelligent text processing and computational linguistics, Springer, pp 1–12.
59. Torraco RJ (2016) Writing integrative reviews of the literature: Methods and purposes. Int J Adult Vocat Educ Technol (IJAVET) 7(3):62–70
60. Tsang AS, Prendergast G (2009) Is a "star" worth a thousand words?: the interplay between product-review texts and rating valences. Eur J Mark 43(11/12):1269–1280. https://doi.org/10.1108/03090560910989876
61. Upadhyaya B, Zou Y, Keivanloo I, Ng J (2014) Quality of experience: What end-users say about web services. In: 2014 IEEE international conference on web services (ICWS), IEEE, pp 57–64
62. Vu PM, Nguyen TT, Pham HV, Nguyen TT (2014) Mining user opinions in mobile app reviews: a keyword-based approach (T). In: 30th IEEE/ACM international conference on automated software engineering (ASE), pp 749–759. https://doi.org/10.1109/ASE.2015.85
63. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis, In Conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, pp 347–354.
64. Xiang Z, Schwartz Z, Gerdes JH Jr, Uysal M (2014) What can big data and text analytics tell us about hotel guest experience and satisfaction? Int J Hosp Manag 44:120–130
65. Xiao S, Wei CP, Dong M (2015) Crowd intelligence: analyzing online product reviews for preference measurement. Inf Manag 53(2):169–182
66. Xu Y, Yu Q, Lam W, Lin T (2017) Exploiting interactions of review text, hidden user communities and item groups, and time for collaborative filtering. Knowl Inf Syst 52:221–254. https://doi.org/10.1007/s10115-016-1005-1
67. Yu D, Mu Y, Jin Y (2017) Rating prediction using review texts with underlying sentiments. Inf Process Lett 117:10–18. https://doi.org/10.1016/j.ipl.2016.08.002
68. Zhu L, Lin Y, Cheng M (2019) Sentiment and guest satisfaction with peer-to-peer accommodation: when are online ratings more trustworthy? Int J Hosp Manag 86:102369. https://doi.org/10.1016/j.ijhm.2019.102369

Publisher's Note