Hägele, Lukas; Klier, Mathias; Moestue, Lars; Obermeier, Andreas

Aspect-based currency of customer reviews: A novel probability-based metric to pave the way for data quality-aware decision-making

Electronic Markets

Provided in Cooperation with:
Springer Nature

Suggested Citation: Hägele, Lukas; Klier, Mathias; Moestue, Lars; Obermeier, Andreas (2025) : Aspect-based currency of customer reviews: A novel probability-based metric to pave the way for data quality-aware decision-making, Electronic Markets, ISSN 1422-8890, Springer, Berlin, Heidelberg, Vol. 35, Iss. 1,
https://doi.org/10.1007/s12525-025-00760-4

Electronic Markets (2025) 35:10
https://doi.org/10.1007/s12525-025-00760-4

RESEARCH PAPER

Aspect-based currency of customer reviews: A novel probability-based metric to pave the way for data quality-aware decision-making

Lukas Hägele[1] · Mathias Klier[1] · Lars Moestue[1] · Andreas Obermeier[1]

Received: 23 May 2024 / Accepted: 17 January 2025

Abstract

Customer reviews from digital platforms are a vital data resource for recommender and other decision support systems. The performance of these systems is highly dependent on the quality of the underlying data—particularly its currency. Existing metrics for assessing the currency of customer reviews are often based solely on data age. They do not consider that customer reviews can be outdated with respect to one aspect (e.g., guest room after renovation) while still being up-to-date with respect to others (e.g., location). Moreover, they disregard that customer reviews can only become outdated due to state changes of the corresponding item (e.g., renovation), which are associated with uncertainty. We propose a probability-based metric for the aspect-based currency of customer reviews. The values of the metric represent the probability that information in a set of customer reviews is still up-to-date. Our evaluation on a large TripAdvisor dataset shows that the values of the metric are reliable and discriminate well between up-to-date and outdated data, paving the way for data quality-aware decision-making based on customer reviews.

Keywords Data quality · Currency · Customer reviews · Data quality metric

JEL Classification M10

Introduction

Today, a large and growing number of customer reviews are available for all kinds of products and services (i.e., items) on many different digital platforms (Yin et al., 2014). For example, the travel-based platform TripAdvisor boasts over one billion customer reviews covering more than eight million restaurants, hotels, attractions, and experiences (TripAdvisor, 2022). This wealth of information, feedback directly from customers who have experienced these items (Brand et al., 2022; Chen, 2023), makes customer reviews one of the most important data sources in e-commerce (Hung et al., 2024). This is especially true for data-driven decision-making, for example, supported by decision support systems or data analytics methods (Sun et al., 2022; Sysko-Romańczuk et al., 2022). However, all these applications require high-quality data in order to provide viable results (Elgendy et al., 2022; Heinrich et al., 2021). In particular, numerous studies underscore the importance of data currency (i.e., whether data still corresponds to its counterparts in the real world) for the performance of decision support systems and data analytics methods (Abraham et al., 2023; Hägele et al., 2024; Holstein et al., 2023; Hristova, 2014). Consequently, customer reviews need to be up-to-date to strengthen decision-making quality (Hägele et al., 2024; Heinrich & Hristova, 2016; McKinney et al., 2002). However, the creation of customer reviews is largely unmonitored (i.e., with little or no quality oversight), allowing anyone to contribute without major barriers (Dhar & Bose, 2022). Furthermore, the creators of customer reviews often fail to update their reviews (Fitchett & Hoogendoorn, 2019).

Responsible Editor: Judith Gebauer.

✉ Mathias Klier
mathias.klier@uni-ulm.de

Lukas Hägele
lukas.haegele@uni-ulm.de

Lars Moestue
lars.moestue@uni-ulm.de

Andreas Obermeier
andreas.obermeier@uni-ulm.de

[1] Institute of Business Analytics, University of Ulm, Helmholtzstr. 22, 89081 Ulm, Germany

Published online: 04 February 2025

This may be due to a lack of motivation on the part of the authors or a lack of functionality of the review platform to update a review (Jin et al., 2014). Another reason may be that the authors do not recognize when their reviews become outdated (e.g., if a hotel renovates its rooms, a past guest would only become aware of this when revisiting the hotel) (Yakubu & Kwong, 2021). Therefore, on the one hand, maintaining a high level of data quality and especially currency in the context of customer reviews is very challenging. On the other hand, a well-founded method for assessing the currency of customer reviews is needed as the basis for any targeted data quality improvement effort (Heinrich & Klier, 2015; Heinrich et al., 2018).

To address this challenge, the assessment of currency can draw on the fact that customer reviews can only become outdated if the state of the corresponding item in the real world changes. For example, a hotel that renovates its guest rooms changes its state, and customer reviews that describe the rooms as old and run-down (which was true when the reviews were created) become outdated. This example also highlights that state changes can be attributed to certain aspects of the item (e.g., the aspect *guest rooms* in the case of a renovation). In this sense, customer reviews may be up-to-date with respect to certain aspects, while being outdated with respect to others (e.g., the renovation may change the state of the *guest rooms*, while not affecting the state of other aspects such as *food* or *location*). In the following, we refer to such changes as aspect-based state changes, because they change the state of the corresponding item in a particular aspect such that the information associated with that aspect in the customer reviews is outdated. Yet, existing approaches for assessing the currency of customer reviews neglect (aspect-based) state changes and use only simple features such as the time since a customer review was created (i.e., their "age"). While, in general, the age of data can indeed influence the performance of data-driven models and the resulting decisions (Lazaridou et al., 2021; Raza and Ding, 2022), findings regarding customer review-based decision support systems, such as recommender systems, are mixed. Some studies indicate that recommender system performance improves when trained on recent customer reviews only (Verachtert et al., 2022), while other studies suggest that even older reviews can enhance recommender system performance (Zheng & Ip, 2013). These contradictory results highlight that the time since creation alone is not a sufficient indicator of the true currency of customer reviews. In particular, reviews with a low age might be assumed to be up-to-date, regardless of whether there has been a respective state change. Vice versa, older reviews might be considered outdated even if no such state change has occurred. In addition, basing the assessment of currency on age alone does not allow for a fine-grained identification of actually outdated or up-to-date passages (i.e., a whole customer review is declared outdated even if only individual aspects are outdated). For example, a hotel review may still be up-to-date concerning the location but outdated regarding room quality.

Therefore, it is important to assess the currency of customer reviews based on aspect-based state changes. Indeed, it has been shown that considering the currency of customer reviews in an aspect-based manner can lead to significantly better decision-making quality (Hägele et al., 2024). In practice, not assessing (aspect-based) currency is problematic in many respects. For example, a customer who prioritizes room quality might not receive a recommendation for a recently renovated hotel because only a few reviews reflect the renovation. As a result, the user may book a less suitable hotel or none at all, leading to suboptimal outcomes for everyone involved. The hotel loses a potential guest, while the customer misses out on a better experience and may settle for a less favorable one—or none. This, in turn, negatively impacts the customer's perception of the platform, lowering customer satisfaction and potentially harming the platform's reputation.

Therefore, we aim to make a Design Science Research (DSR) contribution by answering the research question "how to design a probability-based metric for assessing currency of customer reviews that accounts for aspect-based state changes." Specifically, we design and evaluate a new method to assess the currency of customer reviews (on an aspect level). Thereby, on the one hand, we provide a methodological contribution to data quality, which "has been a central IS research topic for decades" (Padmanabhan et al., 2022, p. vii) and whose importance is constantly emphasized in the IS literature (cf. e.g., Abraham et al., 2023; Peng et al., 2023). On the other hand, we contribute to the ongoing debate on the benefits of customer reviews for digital platforms as one of the most prominent examples of user-generated content and a core component of digital platforms (cf. e.g., Brand et al., 2022; Wrabel et al., 2022). Indeed, customer reviews often serve as a data basis for recommender systems, where it has already been shown that high data quality is important to achieve high performance (Heinrich et al., 2021).

Our proposed novel metric for the aspect-based currency of customer reviews based on the identification of aspect-based state changes. Hereby, it is not known with certainty if, when, and how often aspect-based state changes occur, as explicit data in this regard is typically not available. We argue that the principles and the knowledge base of probability theory are adequate and valuable, providing well-founded methods for describing and analyzing such situations under uncertainty. Thus, we develop our aspect-based, state change-driven metric based on probability theory. The metric values can be defined unambiguously and are interpretable as probabilities (Klier et al., 2021). Moreover, they can support decision-making, for example, by integrating

them into expected value calculus (Heinrich & Klier, 2015). We demonstrate the applicability of our metric and evaluate its values in terms of reliability and ability to discriminate between up-to-date and outdated customer reviews. Focusing on the *guest rooms* aspect, we use a large real-world dataset of over three million customer reviews for 1500 hotels from the 50 largest cities in the USA of the platform TripAdvisor. The results of the evaluation show that the metric values are reliable and allow a clear discrimination between up-to-date and outdated customer reviews.

The remainder of the paper is organized following the publication schema proposed by Gregor and Hevner (2013). Specifically, after illustrating the problem context in the next section, we provide an overview of the background and prior works. Then, we develop a novel aspect-based, state change-driven currency metric for customer reviews. We instantiate our metric using a large real-world dataset from TripAdvisor and evaluate the metric values in terms of reliability and ability to discriminate between up-to-date and outdated instances. Afterwards, we discuss implications for theory and practice, reflect on limitations, and provide an outlook for future research. Finally, we conclude with a brief summary.

Problem context

A customer review is a textual and/or numerical evaluation of an item by someone who has previously purchased, visited, or used the item usually published on a digital platform (Biswas et al., 2022). These evaluations help mitigate the information asymmetry between item providers and customers (Hossain et al., 2022). Anyone can create customer reviews without major barriers, and there are few restrictions on the style or format of the text (Mudambi & Schuff, 2010). This results in a large number of customer reviews with a wealth of information about the corresponding item, which can serve as a valuable data resource for data-driven decision-making (Shen et al., 2015; Sun et al., 2022; Sysko-Romańczuk et al., 2022). A prime example are recommender systems, which address the problem of customer information overload on e-commerce platforms (Lowin et al., 2023), leading to increased revenues (Bawack et al., 2022) and higher customer satisfaction (Hanafizadeh et al., 2021; Lu et al., 2015). In addition, platforms and item providers can benefit from the information contained in customer reviews—for example, by using data analytics methods such as machine learning (Bawack et al., 2022). Thereby, the currency of customer reviews is of great importance, as outdated data can lead to incorrect conclusions (Bayraktarov et al., 2019; Sadiq & Indulska, 2017). In particular, decision support systems such as recommender systems and data analytics tasks (e.g., application of machine learning algorithms) that use customer reviews as input data perform poorly if the data is not up-to-date (Birkbeck et al., 2022; Ferencek & Kljajić Borštnar, 2020; Lu et al., 2015). Thus, to leverage the benefits of customer reviews for recommender systems and data analytics, it is crucial that the information contained in customer reviews reflects the current state of the item in the real world. However, it is challenging to ensure high quality and especially currency, as a manual assessment is not economically feasible due to the sheer number of customer reviews (Paul et al., 2017). We further illustrate our problem context by introducing an example in the form of a set of customer reviews for a hotel with an exemplarily focus on the aspect *guest rooms* (cf. Table 1).

Based on the first three customer reviews, a recommender system would typically not suggest this hotel to a customer with a preference for modern guest rooms. This follows the intuition that the ratings for the rooms are low and the respective polarities in the review texts are mostly negative, informing about run-down rooms with old furniture. Since January, however, the reviews show a different picture. In fact, these reviews indicate a recent renovation. As a result, the rooms are no longer run-down, but stylish and attractive. This is reflected in the latest room ratings and texts only. Therefore, the customer reviews created before the renovation may hinder informed decision-making and lead to reduced performance in data analytics tasks that reflect outdated facts (i.e., they describe an outdated state of the rooms that is no longer up-to-date as the renovation has changed the state of the rooms).

Therefore, assessing the currency of customer reviews is of particular importance. In order to base this assessment on

Table 1 Illustrative example of the problem context

Review No	Timestamp	Excerpt from customer review	Room rating
Review 1	23.07.2023	"[…] but the rooms are outdated […]"	2
Review 2	28.07.2023	"[…] my room was old with broken furnishing […]"	2
Review 3	11.08.2023	"[…] rooms are comfy but need refurbishment."	3
…	…	…	…
Review 94	02.04.2024	"Modern and stylish room design!"	4
Review 95	18.04.2024	"[…] awesome newly renovated rooms […]"	5
Review 96	23.04.2024	"[…] the rooms were nice […]"	4

a well-founded definition of currency in our context, we draw on the general interpretation of currency as one dimension (along with, for example, accuracy and completeness) of the multidimensional construct of data quality (Chengalur-Smith et al., 1999; Lee et al., 2002; Redman, 1997). Currency measures whether data in an information system is up-to-date, i.e., whether the data in the information system still corresponds to its counterparts in the real world (Heinrich & Klier, 2015; Nelson et al., 2005; Redman, 1997). In this vein, we define customer reviews as up-to-date if the information contained still reflects the current state of the corresponding product or service in the real world. More precisely, customer reviews are outdated with respect to an aspect if the state of this aspect has changed in the real world and vice versa. In the illustrative example, it is obvious that the given set of customer reviews is no longer up-to-date with respect to the aspect *guest rooms* since it contains reviews that do not reflect the current state of the rooms. Nonetheless, the set of reviews may still be up-to-date with respect to other aspects mentioned, such as the *location* of the hotel or the *food* in the hotel's restaurant, if no state changes have occurred with respect to these aspects. Thus, we aim for an aspect-based assessment of the currency of customer reviews, since customer reviews may reflect the current state of the real world with respect to one aspect while being outdated with respect to another aspect. Indeed, identifying the occurrence of aspect-based state changes is crucial for assessing the aspect-based currency of customer reviews. However, it is usually not known if, when, or how often such aspect-based state changes occur, since respective explicit data is typically missing. Moreover, such state changes cannot be expected to occur with predictable regularity. Thus, the assessment of currency is tied to the uncertainty associated with the occurrence of aspect-based state changes. Situations that involve uncertainty can be effectively analyzed using methods that rely on the principles and knowledge base of probability theory (Grimaldi et al., 2023). Additionally, data quality metric values representing probabilities offer numerous benefits (Heinrich & Klier, 2015). For example, they have a measurable unit, are interval-scaled, and can be used for calculating expected values. Therefore, our goal in developing a metric for the aspect-based currency of customer reviews is to base it on probability theory and thus provide an indication rather than a verified (binary) statement under certainty. In particular, the values of our metric are intended to represent the probability that the information associated with an aspect in a set of customer reviews is still up-to-date.

Background and related work

In this section, we describe prior prescriptive knowledge and existing artifacts in the context of data quality and currency for unstructured data and especially customer reviews. Indeed, there have been significant contributions to the assessment of data quality for structured data (Batini et al., 2011; Lee et al., 2002; Pipino et al., 2002) and some initial efforts for unstructured data (Immonen et al., 2015; Kiefer, 2016, 2019). However, in the context of customer reviews, there is still a lack of research on assessing data quality in general and currency in particular. While approaches for assessing the currency of structured data are not directly applicable to customer reviews due to their unstructured nature, they can still provide valuable starting points for developing respective metrics. Against this background, in this section, we first discuss works regarding the currency of both structured and unstructured data. In particular, we focus on ideas that can serve as starting points for developing a metric for currency of customer reviews. We then provide an overview of existing works on assessing the data quality of customer reviews in general and currency in particular.

Regarding the assessment of currency of structured data, two of the most notable contributions have been made by Ballou et al. (1998) and Even et al. (2010). They model currency (referred to as timeliness by the authors) based on the age (at the instant of assessing currency), a given shelf life of a structured data attribute, and a sensitivity parameter to adapt the metric to the context of the application. To overcome the weakness that not all attributes have a predetermined shelf life, in recent years, probability-based metrics have emerged as a promising avenue for measuring currency. For example, based on the assumption that the shelf life follows an exponential distribution (Heinrich & Klier, 2011) or by modeling currency with Markov chains (Wechsler & Even, 2012). Another approach is to incorporate conditional expectations and additional metadata into the calculation of the metric values (Heinrich & Klier, 2015). Probability-based metrics have the advantage that their values are interval-scaled and interpretable (Heinrich & Klier, 2009, 2015). Despite their advantages and their potential to automatically assess the currency of structured data, all these approaches are defined for structured data with separate attributes that are not available in unstructured data, such as customer reviews. As a result, they cannot be directly applied to customer reviews. Nevertheless, the concept of an automated, probability-based metric that provides interpretable results seems promising and may be adapted in the context of customer reviews as well.

The literature also provides initial contributions regarding the assessment of the currency of unstructured data (Batini & Scannapieco, 2016; Firmani et al., 2016; Hao et al., 2020; Shah et al., 2015; Zhu & Gauch, 2000). For example, when assessing the currency of big data environments (Firmani et al., 2016), knowledge bases (Shah et al., 2015), and websites (Zhu & Gauch, 2000), currency is often represented by the time since the last update (Batini & Scannapieco, 2016). The advantage of the time since the last update as an

indicator for currency is that it is available for many applications of unstructured data in general and customer reviews in particular. However, in the context of customer reviews containing different aspects, the time since the last update is problematic as a sole currency indicator, since customer reviews can still be up-to-date with respect to certain aspects while being outdated with respect to others. In addition, aspect-based state changes, such as major renovations, do not occur with predictable regularity. Therefore, a metric based on the time since the last update cannot account for the uncertainty of these aspect-based state changes.

In the literature on customer reviews, data quality is often associated with the helpfulness of a review (Almagrabi et al., 2015), e.g., by measuring the proportion of helpful votes it receives (Chua & Banerjee, 2016). As a result, many studies focus on automatically determining the helpfulness of customer reviews and investigating the impact of different features of customer reviews on their helpfulness (Almagrabi et al., 2015). To estimate the helpfulness of customer reviews, some researchers use regression models to calculate a helpfulness score between zero and one (Kim et al., 2006; Lee & Choeh, 2014; Zhang & Varadarajan, 2006), while others use classifiers to determine whether a review is helpful or not (Ghose & Ipeirotis, 2011; Hong et al., 2012; Malik & Hussain, 2017). Although these approaches can accurately predict the helpfulness of customer reviews based on features such as review length, number of spelling errors, and subjectivity scores, they do not account for data quality as a multidimensional construct, nor do they account for currency in particular. In addition, the helpfulness of customer reviews is not differentiated for different aspects of an item. Therefore, approaches that focus on the helpfulness of customer reviews cannot identify quality issues related to specific dimensions, such as currency, nor can they identify data quality issues related to specific aspects of the item.

Despite the importance of the currency of customer reviews for data-driven decision-making, only very few researchers have addressed the automatic assessment of currency of customer reviews. They also rely on age, either measured in days since the review was created (Meng et al., 2021) or as the number of days between the customer review and the first customer review created, to assess the currency of customer reviews (Chen & Tseng, 2011). However, neither the information that a review was created a certain number of days after the first review nor the information that a customer review was created a certain number of days ago is directly related to the extent to which the respective review is still up-to-date. This is due to the fact that the currency of customer reviews is tied to the occurrence of state changes of the reviewed items, such as hotel renovations, which occur in unregular patterns that are not strictly related to age or time. The inability to account for these state changes renders the aforementioned approaches inappropriate for assessing the currency of customer reviews, as their accuracy is compromised. In addition, it is necessary to assess the currency of customer reviews on an aspect level, since the information in a customer review can be up-to-date with respect to one aspect and outdated with respect to another. However, to the best of our knowledge, there is no work that assesses the currency of customer reviews with respect to individual aspects of the item being reviewed.

In summary, there has been significant progress in assessing the currency of both structured and unstructured data. However, in the context of customer reviews, these approaches face challenges due to their reliance on structured data attributes or the consideration of features that are difficult to define for customer reviews, such as shelf life. While some initial efforts have been made to assess the currency of customer reviews, these approaches are limited in their ability to overcome the challenges of assessing the currency of customer reviews. In particular, they assess currency as a sole function of the age of the customer reviews and thus cannot account for state changes of the corresponding item in the real world. Indeed, they do not make use of the rich information (e.g., (textual) feedback) contained in customer reviews that could indicate state changes. Moreover, they do not focus on assessing the currency of customer reviews with respect to different aspects of the associated item. Overall, this leads to an inaccurate and rather coarse assessment of the currency of customer reviews. To address this research gap, we propose a probability-based metric for assessing the aspect-based currency of customer reviews. The metric is based on the identification of aspect-based state changes using statistical outlier tests and the rich information contained in customer reviews. Thus, it accounts for the uncertainty in the occurrence of aspect-based state changes and provides easily interpretable values that represent the probability that the information contained in customer reviews is still up-to-date with respect to an aspect of the item.

A metric for aspect-based currency of customer reviews

In this section, we describe our proposed artifact: a novel, probability-based metric for the aspect-based currency of customer reviews. Specifically, we first describe the general setting and the basic idea of our metric. On this basis, we develop our metric. Finally, we show how the metric can be instantiated using the Grubbs outlier test.

General setting and basic idea

In the context of customer reviews, aspect-based currency expresses whether the information associated with a

particular aspect contained in a set of customer reviews still expresses the current state of that aspect of the corresponding item in the real world at the instant of assessment. As a result, customer reviews can only become outdated with respect to an aspect if the corresponding item in the real world changes with respect to that aspect. Such changes can be either abrupt or gradual over time. Regarding hotel customer reviews, for example, the state of the aspect *guest rooms* changes abruptly in the case of a major hotel renovation project, while it changes gradually over a longer period in the case of inadequate maintenance or when a long-term renovation program is conducted. We refer to such changes that affect the currency of customer reviews as aspect-based state changes. As customer reviews become outdated if and only if such an aspect-based state change occurs, assessing the currency of customer reviews is tied to identifying such state changes. Thus, we base our metric for aspect-based currency on the identification of aspect-based state changes as the underlying causes of outdated aspects of customer reviews. However, identifying aspect-based state changes is associated with uncertainty because they typically do not occur with predictable regularity. Indeed, it is not known with certainty if, when, and how often aspect-based state changes will occur. Such situations under uncertainty can be described and analyzed using well-founded methods based on the principles and knowledge base of probability theory. Thus, we aim to develop a metric based on probability theory. In this line, the values of our metric represent the probability that no state change has occurred for a reviewed item during the observation period, and thus that the associated customer reviews remained up-to-date during this period. Defining the metric values as probabilities has several advantages (Heinrich & Klier, 2015). They have a concrete unit of measurement, are interval-scaled, and can be integrated into expected values calculus.

Given a set of customer reviews as evaluations of the state of different aspects of an item, it contains evidence as to whether or not an aspect-based state change is likely to have occurred over time, e.g., in the texts and/or ratings regarding the respective aspects of the reviews (Hu & Liu, 2004; Sun et al., 2019). For example, the renovation of a hotel's guest rooms may be reflected in a shift towards more comments about modern new rooms, resulting in a higher proportion of positive and fewer negative impressions regarding the rooms compared to before the renovation. We base our metric on indicators derived from customer reviews and the coexistence of positive and negative user impressions that make such evidence regarding the probability of a state change tangible. Examples of such indicators are the relative frequency of positive and negative comments on the respective aspects in the review texts, or positive and negative aspect-based ratings over time (e.g., on a daily or monthly basis, forming an indicator curve over time). Changes in the respective indicator curve suggest a higher probability of an aspect-based state change. However, since individual customers may have different perceptions depending on their subjective preferences and experiences, the indicator curve is subject to random and undirected noise even in the absence of aspect-based state changes (Dellarocas, 2003; Musto & Dahanayake, 2022). Therefore, we aim to identify changes in the indicator curve that go substantially beyond this random noise. The area under the indicator curve is more resistant to (random) noise (Box et al., 2015; Hyndman & Athanasopoulos, 2021) and allows the identification of changes in the indicator curve (Chatfield & Xing, 2019; Shumway & Stoffer, 2017). Thus, basing the design of our metric on the area under the indicator curve allows modeling the probability of an aspect-based state change while being more resistant to (random) noise. Figure 1 illustrates two (aspect-based) indicator curves for the two different cases regarding (aspect-based) state changes: one without a state change (left side) and one with a state change (right side). In the case of no state change, the area under the indicator curve remains approximately the same over all (equally sized) time steps, with only small fluctuations due to the expected and unavoidable noise. In contrast, in the case of a state change, the area under the indicator curve changes substantially compared to the area under the indicator curve of previous time steps. Such substantial changes constitute outliers of the area under the indicator curve of a time step with respect to previous time steps. Indeed, in a mathematical sense, outliers represent observations significantly different from other observations (Grubbs, 1969; Maddala & Lahiri, 1992). Thus, given a partitioning of the indicator curve into time steps, the probability of an aspect-based state change can be identified with the probability of the existence of an outlier in the area under the indicator curve. To determine these probabilities, there are well-established and sound methods from statistical hypothesis testing for outliers. They provide p-values that can be interpreted as the probability that no outlier is present (Chandola & Kumar, 2009; Hodge & Austin, 2004). Based on these p-values, our metric models the probability that a set of customer reviews is still up-to-date. This is achieved by multiplying the individual p-values of all time steps (where each p-value represents the probability that the area under the indicator curve for that time step is not an outlier and thus no state change has occurred) to calculate the metric value (i.e., the probability that no state change has occurred in any of the time steps and thus the set of reviews is still up-to-date). In this sense, our metric is capable of detecting both abrupt and gradual state changes. While changes in the indicator curve from gradual state changes may not be as distinct as those resulting from an abrupt state change, they still show a change in the indicator curve that differs substantially from the expected and unavoidable random noise. Especially, since they often

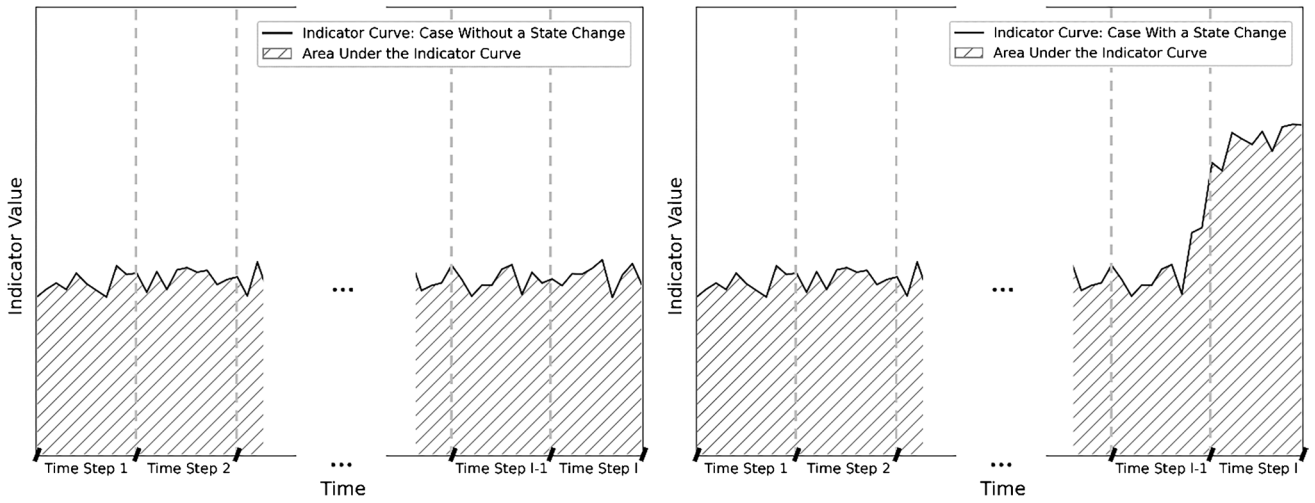

Fig. 1 Indicator curves without state change (left) and with state change (right)

persist across several time steps. Thus, when a gradual state change occurs, the aggregated probability—calculated as the product of the probabilities across individual time steps—that the customer reviews being up-to-date is low (as desired to detect the gradual state change).

To conclude, a set of customer reviews can only become outdated with respect to an aspect due to an aspect-based state change of the corresponding real-world entity. The probability of an aspect-based state change can be identified with the probability of an outlier (i.e., a substantial change) in the area under the indicator curve. Consequently, we define our currency metric as the probability that no outlier (and thus no state change) has occurred regarding the area under the indicator curve at any time step and base its calculation on the *p*-values of a statistical outlier test.

Design of the basic model of the metric

To design our metric, we model the probability that a set of customer reviews R for a particular item within the time period $[t_0;t_1]$ is still up-to-date with respect to an aspect a of the corresponding item at the instant of assessment t_1. Customer reviews can only become outdated with respect to an aspect a if a state change occurs that changes the state of aspect a of the corresponding item in the real world. Thus, the probability that the set of customer reviews is still up-to-date at the instant of assessment t_1 corresponds to the probability that no respective state change has occurred in the time period $[t_0;t_1]$. Given a partitioning of the time period $[t_0;t_1]$ into (equally-sized) time steps $[s_0;s_1], \ldots, [s_{l-1};s_l]$ (with $t_0 = s_0 < \cdots < s_l = t_1$), the probability that no state change occurred in $[t_0;t_1]$ is equivalent to the probability that no state change occurred in any of the time steps $[s_0;s_1], \ldots, [s_{l-1};s_l]$. To avoid a possible loss of valuable information from previous time steps when assessing the probability for a subsequent time step, we formalize the probabilities for subsequent time steps by means of conditional probabilities. This consideration of multiple time steps and the path describing our probability of interest is illustrated by the tree diagram shown in Fig. 2.

Here, O_k for $k = 1, \ldots, l$ denotes the (probability-theoretic) event that a state change with respect to aspect a occurred in the k-th time step $[s_{k-1}, s_k]$. In contrast, $\overline{O}_k$ represents the counter-event that no respective state change occurred in the k-th time step. On this basis, in Eq. (1), we define our metric (i.e., the probability that no state change has occurred in $[t_0;t_1]$ and thus the probability that the set of customer reviews R is still up-to-date) by multiplying all conditional probabilities along the path with no state change occurring in any of the time steps:

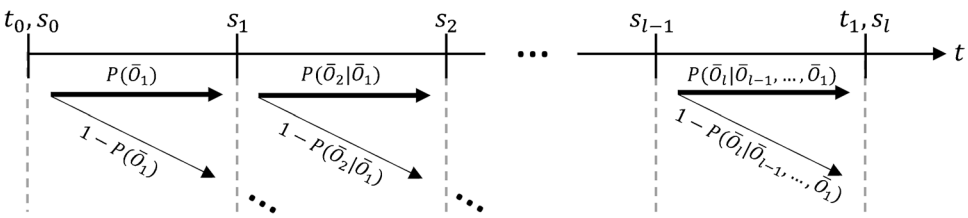

Fig. 2 Tree diagram highlighting the path of probabilities that no outlier occurs

$$Q_{CURR}(R,t_0,t_1) = P\left(\overline{O}_1\right) \cdot P\left(\overline{O}_2|\overline{O}_1\right) \cdot \cdots \cdot P\left(\overline{O}_l|\overline{O}_{l-1},\ldots,\overline{O}_1\right) \quad (1)$$

To determine the probabilities in Eq. (1), we employ that the probability that no aspect-based state change occurred in the k-th time step (i.e., in $[s_{k-1};s_k]$) corresponds to the probability that A_k is not an outlier with respect to the areas under the indicator curve $A_1,\ldots,A_{k-1}$ in the previous time steps. For the first time step $[s_0;s_1]$, no previous area under the indicator curve is available in $[t_0;t_1]$. Thus, another well-founded method is required to estimate $P(\overline{O}_1)$. For example, it would be possible to use a quality-assured reference time step before t_0.

Statistics and the branch of outlier detection based on hypothesis testing offer a rich set of well-founded methods to support the estimation of the required probabilities (i.e., $P(\overline{O}_1)$ and $P\left(\overline{O}_k|\overline{O}_{k-1},\ldots,\overline{O}_1\right)$ for $k = 2,\ldots,l$) such as the Grubbs test, Dixon's Q test, or Thompson Tau test (Chandola & Kumar, 2009; Hodge & Austin, 2004). In particular, the well-known concept of the p-value in hypothesis testing can be used to derive a mathematically sound indication of whether an outlier is present at a given time step (Hodge & Austin, 2004). Indeed, given the null hypothesis that the value under consideration is not an outlier, the corresponding p-value represents the highest level of significance at which this null hypothesis cannot be rejected. Transferred to our context, the probability that the area under the indicator curve in the k-th time step A_k is not an outlier with respect to $A_1,\ldots,A_{k-1}$ can be assessed by means of the p-value p_k of the hypothesis test based on the null hypothesis that A_k is not an outlier with respect to $A_1,\ldots,A_{k-1}$, under the condition that no outlier occurred in these time steps (i.e., $p_1 = P(\overline{O}_1)$ and $p_k = P\left(\overline{O}_k|\overline{O}_{k-1},\ldots,\overline{O}_1\right)$) for $k = 2,\ldots,l$). Finally, the value of our metric for the aspect-based currency of a set of customer reviews R, which represents the probability that no aspect-based state change has occurred in any time step in $[t_0;t_1]$, and thus the information associated with the aspect a in R is up-to-date at the instant of assessment t_1, is given by:

$$Q_{CURR}(R,t_0,t_1) = p_1 \cdot p_2 \cdot \cdots \cdot p_l = \prod_{k=1}^{l} p_k \quad (2)$$

This formalization of the metric as the product of an estimated probability p_k per time step $k = 1,\ldots,l$ that the area under the indicator curve does not represent an outlier favors the detection of both abrupt and gradual state changes. In the case of an abrupt state change, the overall probability $Q_{CURR}(R,t_0,t_1)$ that the customer reviews R are up-to-date is estimated to be small because the p-value p_k for the time step in which the abrupt state change occurs is estimated very low (as A_k constitutes a very clear outlier with respect to $A_1,\ldots,A_{k-1}$). Gradual state changes also show low p-values. While not as distinct as in the case of a sudden state change, for gradual state changes, these low values extend across several time steps. Therefore, when a gradual state change occurs, the aggregated probability of the customer reviews being up-to-date (as a product of all p-values) is low (as desired to detect the gradual state change).

Operationalization of the basic model using the Grubbs outlier test

Our metric provides the probability that the information associated with an aspect a in a set of customer reviews R is up-to-date by assessing the probability of occurrence of an aspect-based state change in the respective time period $[t_0;t_1]$. To be able to determine p_1 of the first time step analogously to p_k for the following time steps $k = 2,\ldots,l$, we additionally use a quality-assured reference time step (referred to as time step 0, with area under the curve A_0) before t_0. Equation (2) provides the mathematical definition of our metric based on p-values from statistical hypothesis tests for outliers (i.e., the conditional probabilities p_k for time steps $k = 1,\ldots,l$ that the area A_k is not an outlier with respect to the areas $A_0,\ldots,A_{k-1}$ under the condition that no state change has occurred in the previous time steps). In statistical hypothesis testing, the Grubbs test (Grubbs, 1969; Stefansky, 1972; Thompson, 1935) is a widely used, reliable, robust, and computationally inexpensive choice for detecting outliers (Urvoy & Autrusseau, 2014). Against this background, for the operationalization of our basic model, we use the Grubbs test to determine p_k ($k = 1,\ldots,l$). Here, p_k is the p-value of the Grubbs test, which represents the probability that A_k is not an outlier with respect to the distribution of the areas under the indicator curve in the previous time steps. The test statistic G_k of the Grubbs test is calculated by taking the absolute value of the difference between A_k (i.e., the area under the indicator curve that we are testing for being an outlier) and the expected value $\mu_{A,k}$ of the distribution of the areas under the indicator curve in the previous time steps, divided by its standard deviation $\sigma_{A,k}$:

$$G_k = \frac{|A_k - \mu_{A,k}|}{\sigma_{A,k}} \quad (3)$$

To determine the test statistic, the mean $\mu_{A,k}$ and the standard deviation $\sigma_{A,k}$ of the normal distribution of the areas under the indicator curve in previous time steps are needed. For this purpose, we exploit the fact that the areas under the indicator curve depend on the values of the respective indicator. Under the condition that no state change has occurred in the previous time steps, these indicator values show the typical behavior of a normally distributed random variable (Shumway & Stoffer, 2017), since they are nearly constant with small

fluctuations (i.e., random noise) characterized by their mean μ_k, which represents their central tendency, and their standard deviation σ_k as a measure of the fluctuation. To determine the area under the indicator curve, we use numerical integration methods (Davis & Rabinowitz, 1984), taking advantage of the fact that these methods calculate the area under the indicator curve as a weighted sum of the indicator values (Davis & Rabinowitz, 1984). Thus, the area under the indicator curve (under the condition that no state change has occurred) is calculated as the weighted sum of n_k normally distributed indicator values (where n_k is the total number of indicator values in the previous time steps) which by definition is also normally distributed (DeGroot & Schervish, 2010) with parameters $\mu_{A,k}$ and $\sigma_{A,k}$. Both parameters $\mu_{A,k}$ and $\sigma_{A,k}$ can be calculated in terms of μ_k and σ_k from the distribution $N(\mu_k, \sigma_k^2)$ of the indicator values. Using the trapezoidal rule, a well-known and widely used numerical integration method (Atkinson, 1989; Rahman & Schmeisser, 1990), the parameters are given by $\mu_{A,k} = \mu_k(s_k(n_k - 1))$ and $\sigma_{A,k} = \sigma_k(s_k\sqrt{n_k - 1.5})$ with s_k being a scaling factor. The scaling factor accounts for the differences in width between the k-th time step and the previous time steps and the resulting difference in magnitude between A_k and the areas given by $N(\mu_{A,k}, \sigma_{A,k}^2)$. On this basis, $\mu_{A,k}$ and $\sigma_{A,k}$ are determined to calculate the test statistic of the Grubbs test G_k to assess p_k. The dependencies from all previous time steps (i.e., conditional probabilities) are taken into account by calculating $\mu_{A,k}$ and $\sigma_{A,k}$ based on the indicator values (and their μ_k and σ_k) from all previous time steps $0, ..., k-1$. The null hypothesis that A_k is not an outlier would be rejected at the significance level α_k if G_k was greater than a critical value Z_{α_k,n_k}. This critical value Z_{α_k,n_k} depends on the significance level α_k and the number n_k of indicator values and is calculated as follows:

$$Z_{\alpha_k,n_k} = \frac{n_k - 1}{\sqrt{n_k}} \sqrt{\frac{t^2_{\frac{\alpha_k}{2n_k},n_k-2}}{n_k - 2 + t^2_{\frac{\alpha_k}{2n_k},n_k-2}}} \tag{4}$$

where $t_{\frac{\alpha_k}{2n_k},n_k-2}$ denotes the upper critical value at the significance level $\alpha_k/2n_k$ for a t-distribution with $n_k - 2$ degrees of freedom (with n_k required to be greater than or equal to three (Grubbs, 1969)). The p-value constitutes the highest significance level α_k at which the null hypothesis cannot be rejected (i.e., the test statistic G_k is smaller than the critical value Z_{α_k,n_k}). The critical value Z_{α_k,n_k} decreases as the significance level α_k increases (Grubbs & Beck, 1972) and thus maximizing α_k is equivalent to minimizing the critical value Z_{α_k,n_k}. Using these relationships, the p-value p_k can be assessed in terms of the solution to the optimization problem of searching the significance level α_k for which the critical value Z_{α_k,n_k} reaches its minimum (i.e., α_k is maximal) under the condition that the null hypothesis cannot be rejected (i.e., the test statistic G_k is smaller than the critical value Z_{α_k,n_k}):

$$p_k = \underset{\alpha_k}{\operatorname{argmin}}\left(Z_{\alpha_k,n_k} | G_k \leq Z_{\alpha_k,n_k}\right) \tag{5}$$

Consequently, using Eq. (5) in our metric given by Eq. (2), the value of our metric for the Grubbs test operationalization described in this section is given by:

$$Q_{CURR}(R, t_0, t_1) = \prod_{k=1}^{l} p_k = \prod_{k=1}^{l} \underset{\alpha_k}{\operatorname{argmin}}\left(Z_{\alpha_k,n_k} | G_k \leq Z_{\alpha_k,n_k}\right) \tag{6}$$

Note that our metric can be applied in a fully unsupervised and automated manner, as there is no need for manual intervention, making it particularly suitable for large datasets containing many customer reviews.

Demonstration and evaluation

In this section, we demonstrate and evaluate the practical applicability and utility of our probability-based metric for the aspect-based currency of customer reviews. First, we present the selected case of hotel customer reviews on TripAdvisor. Then, we demonstrate how the metric can be applied to this case. Finally, we evaluate the values of our metric in terms of reliability and discriminative power.

Case selection and dataset

To demonstrate and evaluate our metric, we apply it to a dataset of customer reviews from the review platform TripAdvisor, which is one of the most prominent and important review platforms especially for the tourism and hospitality industry. TripAdvisor, a pioneer in travel-centric online reviews (Litvin et al., 2018), has rapidly evolved into the most recognized platform for sharing experiences about hotels, restaurants, and attractions (Gretzel & Yoo, 2008), accumulating over a billion reviews (Statista, 2023b). Thereby, we specifically focus on hotels, a key sector of the tourism and hospitality industry with a market size of over one trillion USD in 2023 (Statista, 2023a). Indeed, hotels are one of the most important categories on TripAdvisor, with over 1.2 million hotels listed on the platform and thus available for review (Nunkoo et al., 2020). Despite their importance, the quality of hotel customer reviews on TripAdvisor varies and is often criticized (Chua & Banerjee, 2013; Xiang et al., 2018; Xie et al., 2016). Regarding currency, the platform does not provide specific information on whether the information contained in the customer reviews still represents the current state of the corresponding item in the real world, but only the date of creation or the date of the

most recent update of its customer reviews. Therefore, in the case of hotel reviews on TripAdvisor, it seems particularly important and relevant to assess currency in an aspect- and probability-based way, taking into account currency-relevant state changes.

To demonstrate and evaluate our metric, we selected 1500 hotels listed on TripAdvisor. More precisely, we selected the 30 most often reviewed hotels for each of the 50 largest cities (by population) in the USA. This approach ensured a diverse sample, ranging from large, tourist-oriented luxury hotels (e.g., hotels on The Strip in Las Vegas) to smaller, business-oriented hotels (e.g., hotels in Milwaukee's business district). To exemplify the ability of our metric to assess the aspect-based currency of customer reviews, we focus on the aspect *guest rooms* of the hotels, which is among the most important aspects regarding customer satisfaction (Nunkoo et al., 2020). For the demonstration, we assess the currency of the customer reviews created regarding the aspect *guest room* in the time period under consideration, which is between January 1st, 2014, and December 31st, 2019. To this end, we collected a dataset of all 3,356,231 TripAdvisor reviews of the considered hotels in this period and extracted the fine-grained room ratings in order to focus on the aspect *guest room*. The fine-grained room ratings, which are given on a scale of one to five, are readily available on TripAdvisor, as the platform offers a comprehensive rating system that allows reviewers to rate various aspects of a hotel in addition to providing an overall rating.

While our proposed metric can be applied in a fully automated manner—without the need for any manual interventions—a labeled dataset is required for the sole purpose of rigorously evaluating our metric. More precisely, in order to evaluate our metric's capability of assessing the probability of the occurrence of state changes (regarding guest rooms), a gold standard dataset containing verified information on actual state changes is needed. Thus, exclusively for evaluation purposes, we manually labeled the corresponding set of customer reviews for each hotel in our dataset with respect to whether their rooms experienced a state change during the time period under consideration, from January 1st, 2014, to December 31st, 2019. On this basis, it is possible to determine for all reviews whether they are still up-to-date with respect to the aspect *guest room* at the end of the time period under consideration (t_1 = December 31st, 2019). For this purpose, we extensively searched the hotel websites, online news articles related to the hotels, and hotel forums for information on state changes regarding the rooms. We further supported our research for state changes with a Bing AI-powered search for state changes of the hotels' guest rooms to enhance the robustness of our markings. By means of this research, we found abrupt or gradual aspect-based state changes (regarding the aspect *guest room*) for 1017 hotels, while no aspect-based state changes were identified for the remaining 483 hotels. Most state changes were major renovation projects, but we also found updates to hotel maintenance standards, long-term renovation projects, hotel expansions, the introduction of new room types, and hotel-wide smoking bans in rooms. In this sense, our dataset is particularly well suited for a rigor evaluation our metric, as it contains both abrupt and gradual state changes for many instances, while still containing enough instances for which the guest rooms did not experience a state change.

Demonstration of the practical applicability of the metric

Our metric aims to assess the probability of the occurrence of aspect-based state changes by using an indicator that makes evidence regarding these state changes tangible. For our case (hotel customer reviews from TripAdvisor and the aspect *guest rooms*), we thus generate the indicator curve by aggregating these fine-grained room ratings on a three-month level (i.e., January to March, April to June, July to September, and October to December; with the first point of the indicator curve being the average of a hotel's room ratings from January 1st, 2014, and March 31st, 2014). We then partition the time period under consideration (i.e., January 1st, 2014, to December 31st, 2019) into three time steps of two years each. The reference time step to support the calculation of the outlier probability for the first time step is from January 1st, 2012, to December 31st, 2013. Figure 3 shows an example of the indicator curves and time steps for two sample hotels in our dataset. On the left side, we observe an approximately constant indicator curve for a hotel where no state change has occurred. The right side shows the indicator curve of a hotel where that underwent a major renovation around June 2016.

To calculate the metric values, we used the operationalization with the Grubbs test described in the previous section. Specifically, to calculate the area under the curve in the different time steps, we used the trapezoidal rule, which allowed us to calculate the area as a weighted sum of the values of the indicator curve in the corresponding time steps. To calculate $\mu_{A,k}$ and $\sigma_{A,k}$ (i.e., the mean and standard deviation of the distribution of the previous areas under the curve), we used the formulas derived in the previous section (i.e., $\mu_{A,k} = \mu_k(s_k(n_k - 1))$ and $\sigma_{A,k} = \sigma_k(s_k\sqrt{n_k - 1.5})$, where μ_k and σ_k are the mean and standard deviation, and n_k is the number of points in the previous time steps $(0, ..., k-1)$). For example, for the initial time step, the latter is eight, corresponding to the eight points in the reference time step, and increases to 16 for the second time step. The scaling factor s_k, which adjusts for the different widths of the area under the curve, is set to 1 for the first step (due to the equal number of points in the reference phase and the first step), 0.5 for the second step, and 0.33 for the third step. Using these

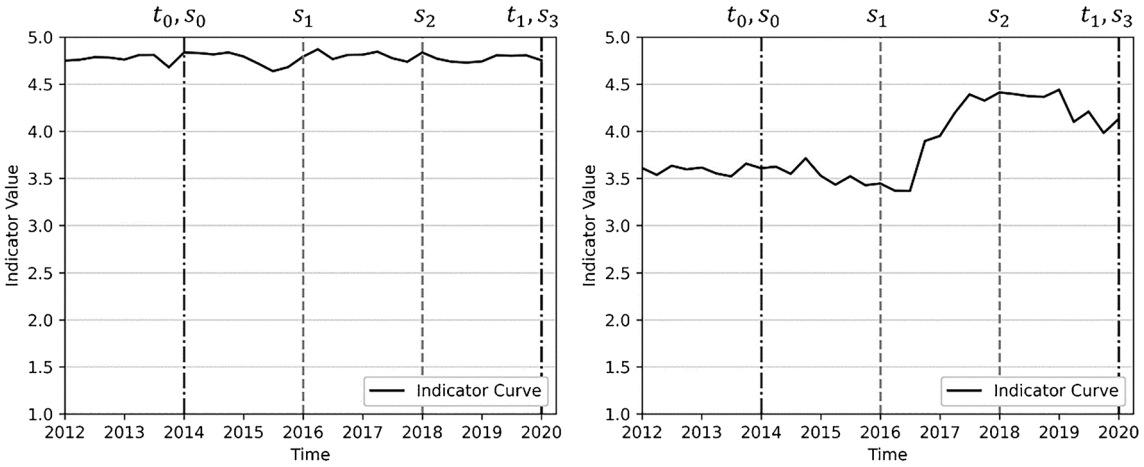

Fig. 3 Example indicator curves for a hotel with no state change (left) and a hotel with a state change (right) for the aspect guest rooms based on the TripAdvisor data

values, we calculate the Grubbs statistic G_k and the critical value Z_{α_k,n_k} (as outlined in Eqs. (4) and (5) of the previous section). This, in turn, allows us to determine p_k by solving the optimization problem $p_k = \underset{\alpha_k}{\operatorname{argmin}}\left(Z_{\alpha_k,n_k} | G_k \leq Z_{\alpha_k,n_k}\right)$ (cf. Eq. (5) in the previous section) using a numerical least-squares optimization approach.

Based on the outlined instantiation, we prototypically implemented our probability-based metric for the aspect-based currency of customer reviews in Python. This allowed us to calculate the probability that the customer reviews between January 1st, 2014, and December 31st, 2019, still expressed the current state of the aspect *guest room* at the end of the time period under consideration (December 31st, 2019) in a fully automated manner for each hotel. Notably, we observe that the metric yields predominantly either very low or very high values, with a large portion (around 60%) of the probabilities falling in the range of 0.20 or lower and about 30% of the probabilities exceeding 0.90. Such a distribution is favorable because it provides the basis for a clear and comprehensive discrimination between up-to-date and outdated instances.

Evaluation of the values of the metric

Our metric estimates the probability that the information associated with a particular aspect in a set of customer reviews still represents the current state of the corresponding item in the real world at the instant of assessment. To analyze whether the proposed metric is able to provide values of high quality, it has to be evaluated whether the estimated probabilities correspond to the actually observed relative frequencies, which can be assessed in terms of reliability (Murphy & Winkler, 1977; Sanders, 1963). In our context, high reliability requires that the metric values in an interval must be approximately equal to the relative frequency of up-to-date instances in that interval. A widely accepted and common way to evaluate reliability is to examine the reliability curve (Bröcker & Smith, 2007). To calculate the points of this curve, the data is sorted into bins according to the metric values. Then, the mean of the metric values ("Mean Metric Value") and the actual relative frequency of instances being up-to-date ("Fraction of Up-to-Date Instances") are calculated and plotted separately for each bin. To obtain enough instances in each bin, the number of bins was set to five. A perfectly reliable probability estimator would be characterized by all points of the corresponding reliability curve lying exactly on the angle bisector, i.e., for each bin, the mean of the metric value and the actual relative frequency of up-to-date instances would be identical. The reliability curve of our metric values is illustrated in the left part of Fig. 4. The results underline that our metric provides reliable values, as the curve follows the angle bisector rather closely. In particular, for very low values, which constitute the majority of cases (about 60%), the relative frequencies of up-to-date instances and the values of our metric coincide almost perfectly. For higher values, our metric slightly overestimates the probability of an instance being up-to-date.

Effective decision-making requires accurate metrics that can discriminate between up-to-date and outdated instances. A widely accepted approach for evaluating the discriminative ability of probability-based metrics is to compute the area under the curve (AUC) of the receiver operator characteristics (ROC) curve (Hanley & McNeil, 1982; Hosmer et al., 2013). Specifically, the ROC curve is constructed by plotting the rate of up-to-date instances correctly classified as up-to-date (true positive rate, TPR) of a classifier based on the estimated probabilities derived from the metric against the rate of outdated

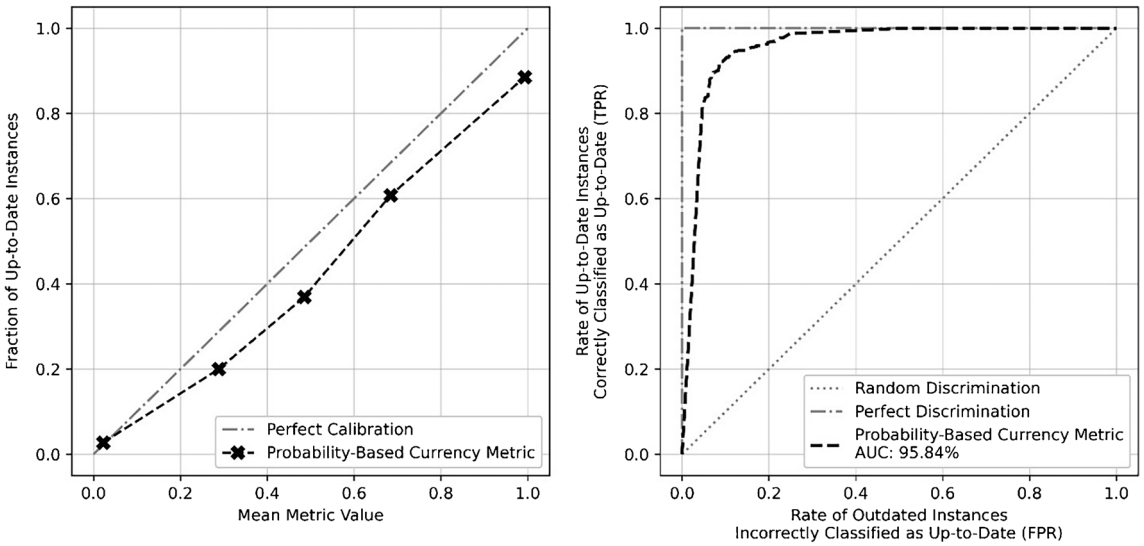

Fig. 4 Reliability curve (left) and ROC curve (right) for our metric values

instances incorrectly classified as up-to-date (false positive rate, FPR) as the classification threshold is varied. In this study, we use the metric values to construct the ROC curve shown in the right part of Fig. 4. Our analysis indicates that the ROC curve closely approximates the curve of perfect discrimination. Furthermore, the AUC value calculated for the proposed metric (95.84%) indicates outstanding discriminative ability (Hosmer et al., 2013). In the following, we further evaluate the discriminative ability of our metric as a binary classifier for the classes *up-to-date* and *outdated* using the common performance measures accuracy, precision, recall, and F1-measure (the harmonic mean of precision and recall). Thereby, we use the probabilities of our metric to classify instances to one of the classes *up-to-date* or *outdated* by assigning them to the class that is most likely; i.e., we use the probability 0.5 as a natural classification threshold for our metric values. As a benchmark to compare the performance of our metric-based binary classifier, we use AutoGluon (Erickson et al., 2020), an open-domain, state-of-the-art automated machine learning framework. AutoGluon automatically employs, optimizes, and combines a diverse set of machine learning models for the given classification task, with the goal of finding the best performing model or ensemble of models. We instantiate AutoGluon in two different ways. In the first case—AutoML (AutoGluon, indicator curve)—AutoGluon uses all values of the indicator curves as input for the classification. In the second case—AutoML (AutoGluon, feature-based)—the indicator curves are first characterized with features such as mean value, standard deviation, minimum and maximum value, and the area under the curve (for each individual time step and the complete indicator curve). We evaluate the performance of AutoGluon using fivefold cross-validation. Since AutoGluon is fully supervised, both versions had access to the labels of the training data in each fold to predict the labels of the test data. For the application of our approach, these labels are not necessary and were collected for reasons of a thorough evaluation only.

For the given dataset, our metric provides very promising results for both the *up-to-date* and the *outdated* class. For all considered performance measures, our metric-based classifier clearly outperforms the automated machine learning framework (cf. Table 2).

Our aspect-based metric correctly classifies 91.26% of the instances and clearly outperforms both AutoGluon instantiations, which achieve an accuracy of 83.33% and 79.66%, respectively. Our metric succeeds in identifying over 90% of both up-to-date and outdated instances (recall), which neither AutoGluon instantiation succeeds in doing for either outdated or up-to-date instances. In addition, our approach is 95.38% and 83.59% correct when predicting outdated and up-to-date instances, respectively (precision), again outperforming both AutoGluon instantiations. Consequently, the F1-measure provides good results for both classes. In particular, a classification performed based on the values of our metric outperforms both AutoGluon instantiations, even with AutoGluon having additional access to the labels of the training data. In conclusion, the results from the evaluation of our metric values confirm that the values of our proposed metric are reliable and can discriminate very well between up-to-date and outdated instances.

Table 2 Performance measures for classification into up-to-date and outdated instances using the instantiation of our proposed metric as well as the two instantiations of AutoGluon

		Accuracy	Recall	Precision	F1-measure
Proposed metric	Outdated	91.26%	91.53%	95.38%	93.41%
	Up-to-date		90.68%	83.59%	86.99%
AutoML (AutoGluon, indicator curve)	Outdated	79.66%	81.28%	87.77%	84.39%
	Up-to-date		76.29%	66.07%	70.81%
AutoML (AutoGluon, feature-based)	Outdated	83.33%	89.34%	85.85%	87.56%
	Up-to-date		71.84%	77.89%	74.75%

Discussion

In this section, we reflect on our probability-based metric for the aspect-based currency of customer reviews. First, we discuss the contributions and implications of our metric for both theory and practice. Then, we elaborate on limitations of our work and propose avenues for future research.

Contributions and implications for theory and practice

Data quality of customer reviews is an important and relevant topic, since the performance of data analytics tasks and decision support systems such as recommender systems depends heavily on the quality and especially the currency of the underlying data. Nevertheless, there are only a few papers that propose currency assessments of customer reviews, all of which measure currency in a simple way depending on the age of the customer reviews. However, it is known from the knowledge base that simple age-based approaches (e.g., for the context of wiki articles, see Heinrich et al., 2023) have weaknesses and can only measure currency to a limited extent. Therefore, in this paper, we designed and evaluated a novel probability-based metric for assessing the aspect-based currency of customer reviews based on the identification of state changes. Our proposed metric contributes to theory and practice in several ways.

The theoretical contributions of our work are twofold. First, our metric is the first that accounts for the process of how customer reviews become outdated, representing an improvement in descriptive knowledge. Specifically, our approach considers state changes, operates at the aspect level, and is based on probability theory. Considering state changes gives credit to the fact that customer reviews can only become outdated if the associated item changes in the real world. Therefore, considering state changes is consistent with the definition of currency as a measure of whether data is up-to-date, i.e., whether the data still corresponds to its counterparts in the real world. However, previous research has focused on the age of the reviews, resulting in inaccurate assessments of currency. Moreover, assessing the currency of customer reviews at the more fine-grained aspect level is beneficial, because even if reviews are outdated with respect to certain aspects, the valuable information they contain with respect to others can still be exploited. Indeed, most customer reviews contain information and ratings regarding different aspects (e.g., location and room condition in the context of hotel reviews). In general, changes in the state of one aspect do not necessarily affect other aspects. Accordingly, customer reviews may be up-to-date with respect to some aspects because they have not changed in the real world, while they are outdated with respect to other aspects. Finally, basing currency assessment on probability theory is required due to the inherent uncertainty associated with the occurrence of aspect-based state changes. This uncertainty arises, since it is not known with certainty if, when, and how often aspect-based state changes occur, because explicit data in this regard is typically not available. In such a case, the principles and knowledge base of probability theory are adequate and valuable, providing well-founded methods for describing and analyzing such situations under uncertainty. Thus, our metric is able to deal with the uncertainty associated with assessing the (aspect-based) currency of customer reviews and its values are interval-scaled and well interpretable as probabilities. As our metric is the first that accounts for the process of how customer reviews become outdated, it constitutes an improvement in the sense of Gregor and Hevner (2013), showing novelty and contributing to the (Ω)-knowledge base.

Second, by proposing our approach based on novel ideas for identifying aspect-based state changes, we contribute to the prescriptive knowledge by improving over existing approaches. Given a set of customer reviews as evaluations of the state of different aspects of an item, it contains evidence as to whether or not an aspect-based state change is likely to have occurred over time, e.g., in the texts and/or ratings regarding the respective aspects of the reviews. We base our novel metric on indicators derived from the customer reviews that make such evidence regarding the probability of a state change tangible. Our aim is to identify changes in the indicator curve that go substantially beyond expected random noise. Therefore, we base our metric on the area under the curve, which is resistant to random noise. In the case of no state change, the area under the curve remains relatively constant, while in the case of a state change, the area under the indicator curve changes substantially. In a mathematical

sense, this means that we are trying to identify outliers in the area under the indicator curve. There are well-established and sound methods from statistical hypothesis testing for outliers to determine the probability of an outlier and thus the probability of an aspect-based state change. This new method contributes to the (Λ)-knowledge base as a level 2 artifact in the sense of Gregor and Hevner (2013). Moreover, our operationalization of the metric with the Grubbs-test contributes to the (Λ)-knowledge base as level 1 artifact. This holds especially as our evaluation shows that the metric values are reliable and can be used for classification into outdated and up-to-date instances. Such a classification based on the metric values even outperforms fully supervised AutoML-approaches requiring additional data. Indeed, our results confirm and extend previous research (Heinrich & Klier, 2011, 2015) in that considering data quality metrics in terms of probabilities is advantageous.

The practical contributions of our work are twofold. First, our metric can help improve decision support systems such as recommender systems that rely on customer reviews. These systems are often hampered by the use of low-quality data, such as outdated customer reviews, which leads to poor performance. Our metric can discriminate well between up-to-date and outdated instances. Therefore, it is possible to use our metric in practical applications to improve the performance of decision support systems by, for instance, eliminating outdated data. For example, in the context of recommender systems, using our metric to filter up-to-date information can improve customer satisfaction and increase revenue for both the platform and the providers of products and services.

Second, our metric can improve the insights of data analytics methods such as machine learning algorithms based on customer reviews. Results based on outdated customer reviews to model the (current) state of the associated item may be invalid. Our metric provides reliable values in the form of probabilities for the aspect-based currency of customer reviews. This information can be incorporated into data-driven decision-making, for example, based on expected value calculus, to increase the validity of the results and foster informed and effective decision-making.

Limitations and future work

While our proposed metric has shown promising results in assessing the aspect-based currency of customer reviews, it is important to acknowledge its limitations and identify potential areas for future research. To demonstrate and evaluate our metric, we focused on a single instantiation with 1500 hotels, using room ratings as an indicator to assess the currency of hotel customer reviews with respect to their guest rooms. Future research could apply the proposed approach with different indicators (e.g., aspect-based sentiments based on the review texts) and compare the results. Similarly, our metric could be applied to customer reviews in other domains (e.g., customer reviews of smartphone apps or restaurants). Even though our metric is able to identify state changes, it cannot specifically identify the type of state change. Future work could employ methods from natural language processing to retrieve the exact causes of such aspect-based state changes from the texts of the customer reviews (i.e., the event that changed the state, such as the renovation of the guest rooms of a hotel). Prior research highlights the positive impact of assessing the currency of customer reviews on decision-making quality (Hägele et al., 2024). Although a study like Hägele et al. (2024) has not yet been conducted for our metric proposed in this work, their initial results with a basic currency metric are compelling. Future research could apply our probability-based, aspect-oriented metric in similar experiments to assess its specific influence on decision quality. Moreover, our metric is the first to be specifically and thoughtfully designed for assessing the aspect-based currency of customer reviews at the individual aspect level. This is especially significant given that customer reviews are often utilized at the aspect level (e.g., Zhang et al., 2014) and can be up-to-date with respect to one aspect while being outdated in others. Nonetheless, potential interconnections between aspects (for instance, improved service quality due to new staff at a hotel might also enhance the room quality through better cleanliness) in customer reviews constitute a highly intriguing avenue for future research in the context of currency metrics for customer reviews in particular. Finally, our metric relies on identifying aspect-based state changes after they have occurred. It may be promising to extend our approach by incorporating ex-ante estimates of the probabilities of such aspect-based state changes in future research. This may be achieved by considering additional information, such as external data indicating aspect-based state changes in the (near) future (e.g., information about planned renovations for a hotel from news articles).

Conclusion

Assessing the currency of customer reviews from digital platforms is an important issue in both theory and practice. In this paper, we propose a novel probability-based metric for assessing the aspect-based currency of customer reviews based on the identification of state changes. Existing metrics are limited in their ability to assess the currency of customer reviews because they are applicable only to structured data, are based only on update frequencies, or model the currency of customer reviews solely by means of simple time intervals. In particular, existing approaches cannot cope with the irregular occurrence of (aspect-based) state changes,

which is the most crucial factor for assessing the currency of customer reviews. To this end, the proposed metric is based on the identification of aspect-based state changes and is formulated in terms of probabilities to account for the uncertainty in their occurrence. We demonstrate the practical applicability of our metric and evaluate its values based on a large real-world dataset of hotel customer reviews from the platform TripAdvisor. The results are promising in that the provided metric values are reliable and allow for a clear discrimination between up-to-date and outdated instances.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The data that support the findings of this study are available from the corresponding author, MK, upon reasonable request.

References

Abraham, R., Schneider, J., & vom Brocke, J. (2023). A taxonomy of data governance decision domains in data marketplaces. *Electronic Markets*, *33*(22). https://doi.org/10.1007/s12525-023-00631-w

Almagrabi, H., Malibari, A., & McNaught, J. (2015). A survey of quality prediction of product reviews. *International Journal of Advanced Computer Science and Applications*, *6*(11), 49–58. https://doi.org/10.14569/IJACSA.2015.061107

Atkinson, K. E. (1989). *An introduction to numerical analysis* (Second Edition). Wiley. http://digitale-objekte.hbz-nrw.de/webclient/DeliveryManager?pid=1641334&custom_att_2=simple_viewer

Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science, 44*(4), 462–484. https://doi.org/10.1287/mnsc.44.4.462

Batini, C., Daniele, B., Federico, C., & Simone, G. (2011). A data quality methodology for heterogeneous data. *International Journal of Database Management Systems, 3*(1), 60–79. https://doi.org/10.5121/ijdms.2011.3105

Batini, C., & Scannapieco, M. (2016). *Data and information quality: Dimensions, principles and technique*. Springer International Publishing. https://link.springer.com/content/pdf/10.1007/978-3-319-24106-7.pdf

Bawack, R. E., Wamba, S. F., Carillo, K. D. A., & Akter, S. (2022). Artificial intelligence in E-Commerce: A bibliometric study and literature review. *Electronic Markets, 32*(1), 297–338. https://doi.org/10.1007/s12525-022-00537-z

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., Possingham, H. P., & Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution*, *6*, Article 239. https://doi.org/10.3389/fevo.2018.00239

Birkbeck, G., Nagle, T., & Sammon, D. (2022). Challenges in research data management practices: A literature analysis. *Journal of Decision Systems, 31*(sup1), 153–167. https://doi.org/10.1080/12460125.2022.2074653

Biswas, B., Sengupta, P., & Ganguly, B. (2022). Your reviews or mine? Exploring the determinants of "perceived helpfulness" of online reviews: A cross-cultural study. *Electronic Markets, 32*(3), 1083–1102. https://doi.org/10.1007/s12525-020-00452-1

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (fifth edition). *Wiley series in probability and statistics*. Wiley. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1061322

Brand, B. M., Kopplin, C. S., & Rausch, T. M. (2022). Cultural differences in processing online customer reviews: Holistic versus analytic thinkers. *Electronic Markets, 32*(3), 1039–1060. https://doi.org/10.1007/s12525-022-00543-1

Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting, 22*(3), 651–661. https://doi.org/10.1175/WAF993.1

Chandola, V., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chatfield, C., & Xing, H. (2019). *The analysis of time series: An introduction with R* (Seventh Edition). *Chapman & Hall/CRC texts in statistical science series*. Chapman and Hall/CRC. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5760899

Chen, C. C., & Tseng, Y.-D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems, 50*(4), 755–768. https://doi.org/10.1016/j.dss.2010.08.023

Chen, J.-S. (2023). Antecedents and outcomes of virtual presence in online shopping: A perspective of SOR (Stimulus-Organism-Response) paradigm. *Electronic Markets*, *33*(1). https://doi.org/10.1007/s12525-023-00674-z

Chengalur-Smith, I. N., Ballou, D. P., & Pazer, H. L. (1999). The impact of data quality information on decision making: An exploratory analysis. *IEEE Transactions on Knowledge and Data Engineering, 11*(6), 853–864. https://doi.org/10.1109/69.824597

Chua, A. Y., & Banerjee, A. (2013). Reliability of reviews on the Internet: The case of TripAdvisor. In *Proceedings of the World Congress on Engineering and Computer Science,* San Francisco, CA.

Chua, A. Y., & Banerjee, S. (2016). Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Computers in Human Behavior, 54*, 547–554. https://doi.org/10.1016/j.chb.2015.08.057

Davis, P. J., & Rabinowitz, P. (1984). *Methods of numerical integration* (Second Edition). Academic Press. http://www.loc.gov/catdir/enhancements/fy0667/2006050288-d.html

DeGroot, M. H., & Schervish, M. J. (2010). *Probability and statistics* (Fourth Edition). Addison-Wesley.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1407–1424. https://doi.org/10.1287/mnsc.49.10.1407.17308

Dhar, S., & Bose, I. (2022). Walking on air or hopping mad? Understanding the impact of emotions, sentiments and reactions on ratings in online customer reviews of mobile apps. *Decision Support Systems, 162*, 113769. https://doi.org/10.1016/j.dss.2022.113769

Elgendy, N., Elragal, A., & Päivärinta, T. (2022). DECAS: A modern data-driven decision theory for big data and analytics. *Journal of Decision Systems, 31*(4), 337–373. https://doi.org/10.1080/12460125.2021.1894674

Erickson, N., Mueller, J., Shirkov, A., Zhang, H, Larroy, P., Li, M., & Smola, A. (2020). *AutoGluon-tabular: Robust and accurate AutoML for structured data*. https://arxiv.org/pdf/2003.06505.pdf

Even, A., Shankaranarayanan, G., & Berger, P. D. (2010). Evaluating a model for cost-effective data quality management in a real-world CRM setting. *Decision Support Systems, 50*(1), 152–163. https://doi.org/10.1016/j.dss.2010.07.011

Ferencek, A., & Kljajić Borštnar, M. (2020). Data quality assessment in product failure prediction models. *Journal of Decision Systems, 29*(sup1), 79–86. https://doi.org/10.1080/12460125.2020.1776927

Firmani, D., Mecella, M., Scannapieco, M., & Batini, C. (2016). On the meaningfulness of "Big Data Quality". *Data Science and Engineering*, *1*(1), 6–20. https://link.springer.com/content/pdf/10.1007/s41019-015-0004-7.pdf

Fitchett, J. M., & Hoogendoorn, G. (2019). Exploring the climate sensitivity of tourists to South Africa through TripAdvisor reviews. *South African Geographical Journal, 101*(1), 91–109. https://doi.org/10.1080/03736245.2018.1541022

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188

Gregor, S., & Hevner, A. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In *Proceedings of the International Conference on Information and Communication Technologies in Tourism 2008,* Innsbruck, Austria.

Grimaldi, D., Sallan, J. M., Arboleda, H., & Sehgal, S. (2023). Rethinking the role of uncertainty and risk in Marketing. *Journal of Decision Systems*, 1–22. https://doi.org/10.1080/12460125.2023.2232570

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1–21. https://doi.org/10.2307/1266761

Grubbs, F. E., & Beck, G. (1972). Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics, 14*(4), 847–854. https://doi.org/10.2307/1267134

Hägele, L. J., Klier, M., Obermeier, A. A., & Sparn, C. (2024). Oldie but goodie – Currency beats data age of customer reviews when it comes to recommender system performance. In *Proceedings of the Forty-Fifth International Conference on Information Systems*, 8, Bangkok, Thailand. https://aisel.aisnet.org/icis2024/data_soc/data_soc/8

Hanafizadeh, P., Barkhordari Firouzabadi, M., & Vu, K. M. (2021). Insight monetization intermediary platform using recommender systems. *Electronic Markets, 31*(2), 269–293. https://doi.org/10.1007/s12525-020-00449-w

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hao, S., Chai, C., Li, G., Tang, N., Wang, N., & Yu, X. (2020). Outdated fact detection in knowledge bases. In *International Conference on Data Engineering,* Dallas, TX.

Heinrich, B., & Hristova, D. (2016). A quantitative approach for modelling the influence of currency of information on decision-making under uncertainty. *Journal of Decision Systems, 25*(1), 16–41. https://doi.org/10.1080/12460125.2015.1080494

Heinrich, B., & Klier, M. (2011). Assessing data currency—A probabilistic approach. *Journal of Information Science, 37*(1), 86–100.

Heinrich, B., & Klier, M. (2015). Metric-based data quality assessment — Developing and evaluating a probability-based currency metric. *Decision Support Systems, 72*, 82–96. https://doi.org/10.1016/j.dss.2015.02.009

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., & Szubartowicz, M. (2021). Data quality in recommender systems: The impact of completeness of item content data on prediction accuracy of recommender systems. *Electronic Markets, 31*(2), 389–409. https://doi.org/10.1007/s12525-019-00366-7

Heinrich, B., & Klier, M. (2009). A novel data quality metric for timeliness considering supplemental data. In *Proceedings of the European Conference on Information Systems,* Verona, Italy.

Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)*, 9(2), Article 12, 1–32.https://doi.org/10.1145/3148238

Heinrich, B., Huber, M. F., Krapf, T., & Schiller, A. (2023). The currency of Wiki Articles – A language model-based approach. In *Proceedings of the Forty-Fourth International Conference on Information Systems*, Hyderabad, India. https://aisel.aisnet.org/icis2023/dab_sc/dab_sc/14

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Holstein, J., Schemmer, M., Jakubik, J., Vössing, M., & Satzger, G. (2023). Sanitizing data for analysis: Designing systems for data understanding. *Electronic Markets*, 33(52). https://doi.org/10.1007/s12525-023-00677-w

Hong, Y., Lu, J [Jun], Yao, J., Zhu, Q., & Zhou, G. (2012). What reviews are satisfactory: Novel features for automatic helpfulness voting. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,* Portland, OR.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Third Edition). *Wiley series in probability and statistics*. John Wiley & Sons.

Hossain, M. A., Akter, S., & Rahman, S. (2022). Customer behavior of online group buying: An investigation using the transaction cost economics theory perspective. *Electronic Markets*, *32*(3), 1447–1461. https://doi.org/10.1007/s12525-021-00479-y

Hristova, D. (2014). Considering currency in decision trees in the context of big data. In *Proceedings of the Thirty-Fifth International Conference on Information Systems*, Auckland, New Zealand. https://aisel.aisnet.org/icis2014/proceedings/DecisionAnalytics/1/

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining,* Seattle, WA.

Hung, H.-Y., Hu, Y., Lee, N., & Tsai, H.-T. (2024). Exploring online consumer review-management response dynamics: A heuristic-systematic perspective. *Decision Support Systems, 177*, 114087. https://doi.org/10.1016/j.dss.2023.114087

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (Third Edition). OTexts. https://otexts.com/fpp2/

Immonen, A., Paakkonen, P., & Ovaska, E. (2015). Evaluating the quality of social media data in big data architecture. *IEEE Access, 3*, 2028–2043. https://doi.org/10.1109/access.2015.2490723

Jin, L., Hu, B., & He, Y. (2014). The recent versus the out-dated: An experimental examination of the time-variant effects of online consumer reviews. *Journal of Retailing, 90*(4), 552–566. https://doi.org/10.1016/j.jretai.2014.05.002

Kiefer, C. (2016). Assessing the quality of unstructured data: An initial overview. In *LWDA 2016 Proceedings,* Potsdam, Germany. http://ceur-ws.org/vol-1670/paper-25.pdf

Kiefer, C. (2019). Quality indicators for text data. In *BTW 2019 - Workshopband,* Rostock, Germany. https://dl.gi.de/handle/20.500.12116/21801

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Sydney, Australia.

Klier, M., Moestue, L., Obermeier, A., & Widmann, T. (2021). Event-driven assessment of currency of Wiki Articles: A novel probability-based metric. In *Proceedings of the Forty-Second International Conference on Information Systems,* Austin, TX. https://aisel.aisnet.org/icis2021/data_analytics/data_analytics/14

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A, Teri, T., de Masson d'Autume, C., Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S., & Blunsom, P. (2021). Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems 34*, A virtual Event.

Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications, 41*(6), 3041–3046. https://doi.org/10.1016/j.eswa.2013.10.034

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management, 40*(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2018). A retrospective view of electronic word-of-mouth in hospitality and tourism management. *International Journal of Contemporary Hospitality Management, 30*(1), 313–325. https://doi.org/10.1108/IJCHM-08-2016-0461

Lowin, M., Oz, D., Somech, I., Zahn, M. von, Hinz, O., & Reichman, S. (2023). Designing profit-maximizing recommender systems in E-commerce: An experimental analysis. *SSRN Electronic Journal.* Advance online publication.https://doi.org/10.2139/ssrn.4553876

Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems, 74*, 12–32. https://doi.org/10.1016/j.dss.2015.03.008

Maddala, G. S., & Lahiri, K. (1992). *Introduction to econometrics*. Macmillan.

Malik, M., & Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior, 73*, 290–302. https://doi.org/10.1016/j.chb.2017.03.053

McKinney, V., Yoon, K., & Zahedi, F. (2002). The measurement of web-customer satisfaction: An expectation and disconfirmation approach. *Information Systems Research, 13*(3), 296–315. https://doi.org/10.1287/isre.13.3.296.76

Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2021). What makes an online review more helpful: An interpretation framework using XGBoost and SHAP values. *Journal of Theoretical and Applied Electronic Commerce Research, 16*(3), 466–490. https://doi.org/10.3390/jtaer16030029

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185–200. https://doi.org/10.2307/20721420

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series c: Applied Statistics, 26*(1), 41–47. https://doi.org/10.2307/2346866

Musto, J., & Dahanayake, A. (2022). Quality characteristics for user-generated content. In M. Tropmann-Frick, B. Thalheim, H. Jaakkola, Y. Kiyoki, & N. Yoshida (Eds.), *Frontiers in artificial intelligence and applications, 343. Information modelling and knowledge bases XXXIII* (pp. 244–263). IOS Press. https://doi.org/10.3233/FAIA210490

Nelson, R. R., Wixom, B. H., & Todd, P. R. (2005). Antecedents of information and system quality: An empirical examination within the context of data warehousing. *Journal of Management Information Systems, 21*(4), 199–235. https://doi.org/10.1080/07421222.2005.11045823

Nunkoo, R., Teeroovengadum, V., Ringle, C. M., & Sunnassee, V. (2020). Service quality and customer satisfaction: The moderating effects of hotel star rating. *International Journal of Hospitality Management, 91*, 102414. https://doi.org/10.1016/j.ijhm.2019.102414

Padmanabhan, B., Fang, X., Sahoo, N., & Burton-Jones, A. (2022). Machine learning in information systems research. *MIS Quarterly, 46*(1), iii–xix.

Paul, D., Sarkar, S., Chelliah, M., Kalyan, C., & Sinai Nadkarni, P. P. (2017). Recommendation of high quality representative reviews in e-commerce. In *RecSys '17: Eleventh ACM Conference on Recommender Systems,* Como, Italy.

Peng, J., Hahn, J., & Huang, K.-W. (2023). Handling missing values in information systems research: A review of methods and assumptions. *Information Systems Research, 34*(1), iii–vii. https://doi.org/10.1287/isre.2022.1104

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218. https://doi.org/10.1145/505248.506010

Rahman, Q. I., & Schmeisser, G. (1990). Characterization of the speed of convergence of the trapezoidal rule. *Numerische Mathematik, 57*(1), 123–138. https://doi.org/10.1007/BF01386402

Raza, S., & Ding, C. (2022). News recommender system: A review of recent progress, challenges, and opportunities. *Artificial Intelligence Review, 55*, 749–800. https://doi.org/10.1007/s10462-021-10043-x

Redman, T. C. (1997). *Data quality for the information age*. Arctech House. https://dl.acm.org/citation.cfm?id=548570

Sadiq, S., & Indulska, M. (2017). Open data: Quality over quantity. *International Journal of Information Management, 37*(3), 150–154. https://doi.org/10.1016/j.ijinfomgt.2017.01.003

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology, 2*(2), 191–201. https://doi.org/10.1175/1520-0450(1963)002%3c0191:OSPF%3e2.0.CO;2

Shah, A. A., Ravana, S. D., Hamid, S., & Ismail, M. A. (2015). Web credibility assessment: Affecting factors and assessment techniques. *Information Research*, *20*(1), 365–391. http://informationr.net/ir/20-1/paper663.html

Shen, W., Hu, Yu, J., & Ulmer, J. R. (2015). Competing for attention: An empirical study of online reviewers' strategic behavior. *MIS Quarterly*, *39*(3), 683–696. https://doi.org/10.25300/MISQ/2015/39.3.08

Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: With R examples* (Fourth Edition). *Springer Texts in Statistics*. Springer. https://doi.org/10.1007/978-3-319-52452-8

Statista. (2023a). *Global hotel and resort industry market size worldwide 2013–2023.* Statista Inc. https://www.statista.com/statistics/1186201/hotel-and-resort-industry-market-size-global/

Statista. (2023b). *Total number of user reviews and ratings on Tripadvisor worldwide from 2014 to 2022 (in millions).* Statista Inc. https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/

Stefansky, W. (1972). Rejecting outliers in factorial designs. *Technometrics, 14*(2), 469–479. https://doi.org/10.2307/1267436

Sun, Q., Niu, J., Yao, Z., & Yan, H. (2019). Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level.

Engineering Applications of Artificial Intelligence, 81, 68–78. https://doi.org/10.1016/j.engappai.2019.02.004

Sun, J., Song, J., Jiang, Y., Liu, Y., & Li, J. (2022). Prick the filter bubble: A novel cross domain recommendation model with adaptive diversity regularization. *Electronic Markets, 32*(1), 101–121. https://doi.org/10.1007/s12525-021-00492-1

Sysko-Romańczuk, S., Zaborek, P., Wróblewska, A., Dąbrowski, J., & Tkachuk, S. (2022). Data modalities, consumer attributes and recommendation performance in the fashion industry. *Electronic Markets, 32*(3), 1279–1292.

Thompson, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics, 6*(4), 214–219. https://doi.org/10.1214/aoms/1177732567

TripAdvisor. (2022). *TripAdvisor investor relations*. https://ir.tripadvisor.com/

Urvoy, M., & Autrusseau, F. (2014). Application of Grubbs test for outliers to the detection of watermarks. In *Workshop on Information Hiding and Multimedia Security,* Salzburg, Austria.

Verachtert, R., Michiels, L., & Goethals, B. (2022). Are we forgetting something? Correctly evaluate a recommender system with an optimal training window. In *Perspectives on the Evaluation of Recommender Systems Workshop co-located with the 16th ACM Conference on Recommender Systems*, Seattle, WA.

Wechsler, A., & Even, A. (2012). Using a Markov-Chain model for assessing accuracy degradation and developing data maintenance policies. In *American Conference on Information Systems,* Seattle, WA.

Wrabel, A., Kupfer, A., & Zimmermann, S. (2022). Being informed or getting the product? How the coexistence of scarcity cues and online consumer reviews affects online purchase decisions. *Business & Information Systems Engineering, 64*(5), 575–592. https://doi.org/10.1007/s12599-022-00772-w

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2018). Assessing reliability of social media data: Lessons from mining TripAdvisor hotel reviews. *Information Technology & Tourism, 18*(1–4), 43–59. https://doi.org/10.1007/s40558-017-0098-z

Xie, K. L., Chen, C., & Wu, S. (2016). Online consumer review factors affecting offline hotel popularity: Evidence from Tripadvisor. *Journal of Travel & Tourism Marketing, 33*(2), 211–223. https://doi.org/10.1080/10548408.2015.1050538

Yakubu, H., & Kwong, C. K. (2021). Forecasting the importance of product attributes using online customer reviews and Google Trends. *Technological Forecasting and Social Change, 171*, 120983. https://doi.org/10.1016/j.techfore.2021.120983

Yin, D., Bond, S. D., & Zhang, H (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, *38*(2), 539–560. https://doi.org/10.25300/misq/2014/38.2.10

Zhang, Z., & Varadarajan, B. (2006). Utility scoring of product reviews. In P. S. Yu (Chair), *the 15th ACM international conference,* Arlington, VA.

Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., & Ma, S. (2014). Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 83–92. https://doi.org/10.1145/2600428.2609579

Zheng, Q., & Ip, H. H. S. (2013). Effectiveness of the data generated on different time in latent factor model. In *Proceedings of the 7th ACM conference on Recommender systems*, Hong Kong, China. https://doi.org/10.1145/2507157.2507202

Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Athens, Greece.