Journal of Retailing and Consumer Services 85 (2025) 104278

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

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

Measuring technology acceptance over time using transfer models based on online customer reviews

Daniel Baier *, Andreas Karasenko, Alexandra Rese

Chair of Marketing & Innovation, University of Bayreuth, Universitätsstraße 30, 95447, Bayreuth, Germany

ARTICLE INFO

Handling Editor: Prof. H. Timmermans

Keywords:
Online customer reviews
Technology acceptance
Transfer models
LLMs (large language models)
Transformer architecture
Generative artificial intelligence chatbots
ChatGPT

ABSTRACT

Online customer reviews (OCRs) are user-generated, semi-formal evaluations of products, services, or technologies. They usually consist of a timestamp, a star rating, and, in many cases, a comment that reflects perceived strengths and weaknesses. OCRs are easily accessible in large numbers on the Internet – for example, through app stores, electronic marketplaces, online shops, and review websites. This paper presents new transfer models to predict technology acceptance and its determinants from OCRs. We train, test, and validate these prediction models using large OCR samples and corresponding observed construct ratings by human experts and generative artificial intelligence chatbots as well as estimated ratings from a traditional customer survey. From a management perspective, the new approach enhances former technology acceptance measurement since we use OCRs as a basis for prediction and discuss the evolution of acceptance over time.

1. Introduction

For quite some time, online customer reviews (OCRs) have been viewed as a rich source of knowledge for firms (Balasubramanian and Mahajan, 2001; Decker and Trusov, 2010; Yang et al., 2019). They provide necessary customer feedback (e.g., questions, concerns, and complaints), deliver feedback in communication (Vermeer et al., 2019; Yang et al., 2019), and help to improve reviewed products, services, or technologies (see Ye et al., 2019). Various methodological approaches have been proposed for these purposes. Popular approaches include word clouds that visualize frequently discussed words or n-grams (see, e.g., Kim et al., 2022), Latent Dirichlet Allocation that distills popular topics from customer segments (Büschken and Allenby, 2016, 2020), and sentiment analysis combined with regression analysis to identify the strengths and weaknesses perceived by customers (see, e.g., Decker and Trusov, 2010; Yang et al., 2019; Hartmann et al., 2023).

However, up to now, few attempts have been made to augment or replace the standard approach to improving new or modified products, services, and technologies that is still widespread (see, Blut et al., 2022) – namely, customer surveys based on the technology acceptance model (TAM) by Davis (1989) or its extensions. Rese et al. (2014) made one of these rare attempts. They developed TAM construct dictionaries that gave a low (or high) construct score for words and n-grams (i.e. a group of n consecutive words like “not easy”), similar to the lexicon-based approach in sentiment analysis. Moreover, they used these dictionaries to predict construct scores for other OCRs based on the contained words or n-grams. Rese et al. (2014) demonstrated that this approach provides similar results to conducting a traditional TAM customer survey with multi-item scales. Although their approach required considerable effort in pre-processing the reviews (text cleaning and word stemming) and has only produced comparably low path coefficients and predictive validity, they came to the moderate conclusion that “it seems that data collection via dedicated customer surveys can be replaced – with some reservations – by the analysis of publicly available (real) online reviews” (Rese et al., 2014, p. 869). Follow-up investigations either checked for the replicability of the results and the quality of the measurement scales (Rese et al., 2017) or tried to enhance semi-automated text analysis with other dictionaries that reflect the TAM constructs (Schreiber, 2020). Again, the studies showed that the improvement proposals derived for products, services, or technologies were useful from a managerial perspective.

In the meantime, the methodologies for analyzing natural language comments have made enormous progress. Large language models (LLMs) and transfer models, such as BERT (Devlin et al., 2018), based on Google’s well-known Transformer architecture (Vaswani et al., 2017), are openly available and can be fine-tuned for multilingual sentiment

This article is part of a special issue entitled: Technology in Retailing published in Journal of Retailing and Consumer Services.

* Corresponding author.
E-mail addresses: daniel.baier@uni-bayreuth.de (D. Baier), andreas.karasenko@uni-bayreuth.de (A. Karasenko), alexandra.rese@uni-bayreuth.de (A. Rese).

https://doi.org/10.1016/j.jretconser.2025.104278
Received 10 October 2024; Received in revised form 26 February 2025; Accepted 27 February 2025
Available online 26 March 2025
0969-6989/

analysis and comparable text categorization tasks based on small datasets for training (see, e.g., Manias et al., 2023). The superiority of these fine-tuned transfer models over (other) machine learning and lexicon-based approaches for sentiment analysis has been demonstrated in many studies (see, e.g., Alantari et al., 2022; Hartmann et al., 2023). Consequently, in this paper, we investigate whether such transfer models can be applied to text categorization and prediction of extended TAM construct scores. We seek to answer the following research question: *Can transfer models complement or replace traditional measurement of extended TAM constructs with customer surveys? Or more specifically: Are extended TAM construct score predictions using OCR-based transfer models valid?*

To answer this research question, large OCR datasets (reviewing apps, such as Baur, Disney+, Facebook, Ikea, Netflix, and Prime), corresponding human expert and generative artificial intelligence chatbot ratings, as well as a related TAM customer survey are used for training, testing, and validation. First, OCRs for one app (the Ikea1 sample with n = 5356 Ikea app OCRs) were independently rated by three advanced master students with respect to six extended TAM constructs. These so-called human experts were made familiar with extended TAM multi-item scales in advance to clarify the intended meaning of the six constructs. They were asked to rate each OCR comment on a 5-point Likert scale (such as item responses in a TAM questionnaire or OCR star ratings). Three generative artificial intelligence chatbots were used to validate these ratings. Then, transfer models for each construct were trained and tested using the OCR comments and their corresponding construct ratings by the human experts. In the second step, the transfer model's extended TAM construct score predictions were validated by comparing them with measurements derived from other OCR datasets (Ikea2 and Ikea3, also with OCRs for the Ikea app). Finally, in the third step, the transfer model was applied to other OCR datasets with promising results.

Our new approach extends the research by Rese et al. (2014, 2017), and Schreiber (2020) in that the lexicon-based approach used there is replaced by Transformer-based transfer models here and that larger samples of OCRs are analyzed. The contribution of this paper is that the new approach outperforms the former and hereby complements or even replaces traditional TAM measurement based on customer surveys successfully. Moreover, as a second contribution of this paper, the new approach predicts TAM construct scores over time.

The paper is structured as follows. We start in Section 2 with a discussion of recent progress and deficits in TAM modeling and OCR analysis. Section 3 discusses LLMs and the Transformer architecture and how they are fine-tuned to become transfer models for related text categorization of multilingual comments. Section 4 describes our datasets and our research methodology in detail. Then, in Section 5, we present the training and testing of our transfer models based on the dataset Ikea1 (OCRs on the Ikea app collected within a specific time frame) and how these models are validated by applying them to datasets Ikea2 and Ikea3. Section 6 describes the application of our transfer models to other OCR datasets. Finally, Section 7 debates the results achieved and the limitations of our methodology, Section 8 includes the theoretical, methodological, and managerial implications, and Section 9 presents the study's conclusions and outlook.

2. Background: Technology acceptance measurement and online customer reviews (OCRs)

2.1. Technology acceptance measurement

Originally developed to explain the acceptance of an information system in an organization, TAM (Technology Acceptance Model by Davis, 1989) and its extensions, such as TAM2 (Venkatesh and Davis, 2000), UTAUT (unified theory of acceptance and use of technology by Venkatesh et al., 2003), and UTAUT2 (Venkatesh et al., 2012) are today applied to various products, services, or technologies in a wide range of contexts (see the overviews by Mortenson and Vidgen, 2016; Blut et al., 2022). Applications range from overall evaluations of products and services for individuals, groups, and organizations (e.g., access to artificial intelligence, databases, computational/storage resources, construction tools, office software, metaverse apps, online games, online shops, portals, platforms, search engines, social networks) to specific evaluations of selected technologies used in products or services (e.g., augmented and virtual reality, chatbots, recommender systems, search functions, speech assistants, and other enabling tools for various purposes). Often the goal is to predict whether a new product (service or technology) would be adopted and to develop proposals for improvement (Mortenson and Vidgen, 2016). Several overviews and meta-analyses (see Legris et al., 2003; Wu et al., 2011; Mortenson and Vidgen, 2016; Blut et al., 2022) confirm that TAM and its extensions are among the most widely cited and applied theories not only in computer science and service management but also in marketing and other disciplines.

Fig. 1 shows the basic TAM, where two specific beliefs – perceived ease of use and perceived usefulness – are assumed to be linked as predictors to two other constructs – attitude toward using and behavioral intention to use. Davis (1989), as well as Davis et al. (1989), defined these two constructs as follows: Perceived usefulness is "the degree to which a person believes that using a particular system would enhance his or her job performance" (Davis, 1989, p. 320), whereas perceived ease of use is "the degree to which a person believes that using a particular system would be free of effort" (Davis, 1989, p. 320), the latter having an additional indirect positive effect on perceived usefulness. Attitude toward using reflects "an individual's positive or negative feelings (evaluative aspect) about performing the target behavior" (Fishbein and Ajzen, 1975, p. 216, Venkatesh et al., 2003, p. 428) whereas behavioral intention to use works as a proxy for user acceptance (Venkatesh et al., 2003). Empirical studies with samples of users responding to questionnaires containing corresponding indicators of these constructs confirmed the proposed positive relationships between the four constructs (see the overview in Blut et al., 2022).

The overall validity of TAM and related models in predicting usage based on the discussed predictors and its usefulness in explaining acceptance of new technology has often been demonstrated. During decades of research on TAM, supplementary constructs were redefined and/or added to extend the model's applicability or its explanatory and predictive power. For example, perceived usefulness was replaced in UTAUT/UTAUT2 by performance expectancy and perceived ease of use by effort expectancy (Venkatesh et al., 2003). Above all, exogenous constructs and variables were integrated as additional predictors (e.g., education, facilitating conditions, habit, and social influence; see overviews, e.g., Blut et al., 2022).

With particular regard to consumer-oriented products, services, or technologies that have a clear voluntary focus (e.g., augmented and virtual reality enrichments in games, social networks, and shops), Olsson et al. (2013) along with Rese et al. (2014, 2017), and Schreiber (2020) established the usefulness of two additional exogenous constructs – perceived informativeness and perceived enjoyment – to reflect complementary utilitarian and hedonic benefits: Perceived informativeness is the degree to which a person believes that the system supports her or him with "relevant and useful … information" for her or his tasks (Rese et al., 2017, p. 5). Perceived enjoyment is defined as "the extent to which the activity of using … is perceived to be enjoyable in its own right" (Davis et al., 1992, p. 1113). This addition to TAM was in line with research in the field of many consumer-oriented and employee-oriented offers – for example, online shops and access to information and transactions (Chen and Tan, 2004; Bruner and Kumar, 2005; Hausman and Siekpe, 2009). In empirical studies, both constructs have been shown to positively influence attitude toward using and behavioral intention to use (Hausman and Siekpe, 2009; Pantano and Naccarato, 2010; Pantano and Servidio, 2012). Nevertheless, they have been proposed as having a positive effect on perceived usefulness in satisfying different individual

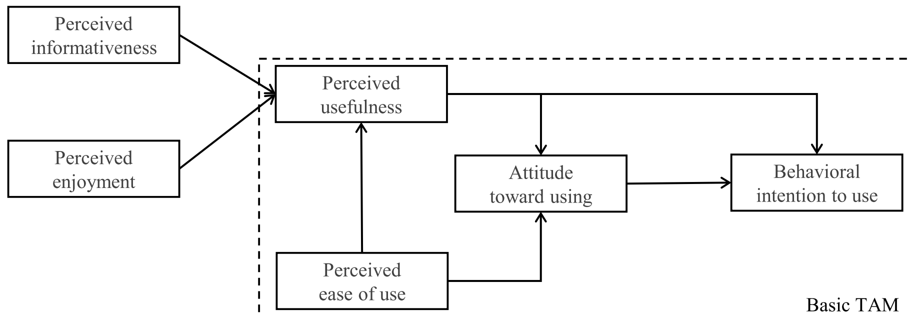

Fig. 1. Basic and extended TAM (Technology Acceptance Model).

needs. The information offered satisfies an individual's informational needs and supports them in their follow-up activities (Chen and Tan, 2004). An offer perceived as entertaining (high perceived enjoyment scores) satisfies individual needs for escapism, diversion, aesthetic enjoyment, or emotional release (Ducoffe, 1996). Fig. 1 summarizes these considerations concerning an extended TAM compared to the basic TAM. The extended TAM is the basis for our empirical investigations, even though we propose later, that our methodology could also be applied to other extensions of TAM/UTAUT. Table 1 presents a generic operationalization of the here-discussed extended TAM constructs, as used by Rese et al. (2014) and follow-up studies (Rese et al., 2017; Schreiber, 2020). The operationalization of the six constructs follows well-known published TAM questionnaires that refer to consumer-oriented technologies (e.g., Ahn et al., 2004, Venkatesh and Davis, 2000, see the discussion in Rese et al., 2014 for details). Typically, the information system's name under study (e.g., Auto Bild app, Ikea app, Mister Spex app, Ray Ban app as used in Rese et al., 2017) is included in all items to clarify the respondent's rating task (app name instead of " …" in Table 1).

However, in addition to the usefulness of TAM and its extensions, the deficits discussed in the literature should be taken into account. Blut et al. (2022) mention in their recent overview and meta-analysis of 1451 articles with 1935 independent samples from 737,112 users across 77 countries, that, despite the widespread impact of TAM and its extensions in computer science, service management, as well as marketing research and practice, the use of the model today has come up against certain limits. The authors link these criticisms to their observation that many important topics – up to now – have not been tackled by the TAM research stream. Consequently, they propose to employ other data collection methods and research designs. That is to say, more observations and qualitative data should be included in the analysis, and longitudinal effects should be investigated (Blut et al., 2022). One proposed solution is to supplement the analysis by focusing on the strengths and weaknesses discussed in the OCR comments and including timestamps.

Table 1
Sample indicators for extended TAM constructs (generic formulations that have to be adapted for a specific product, service, or technology by replacing " …") according to Rese et al. (2014) with 5- or 7-point Likert scales as possible responses for each indicator (from 1 = strongly disagree to 5 = or 7 = strongly agree).

Perceived enjoyment	**Perceived usefulness**
Using … is fun.	For me, …has great value.
… contains nice gimmicks as functions.	… provides beautiful ideas.
It is fun to discover the functions of …	…. is very inspiring in terms of ideas.
… invites you to discover more functions.	… is perfect for keeping the overview.
Perceived informativeness	**Attitude toward using**
… showed the information I expected.	I am positive about …
… provides detailed information.	… is so interesting that you just want to learn more about it.
… provides complete information.	It just makes sense to use …
… provides information that helps.	The use of … is a good idea.
… provides information for comparisons.	Other people should also use …
Perceived ease of use	**Behavioral intention to use**
I found … to be very easy to use.	In the future, I would use … immediately.
… was intuitive to use.	In the future, I would give … priority over other products/services/technologies.
It was easy to learn how to use …	In the future, I would give … priority over other offers of the same company.
Handling the functions of … was easy.	I will recommend using … to my friends.
	I will use … offer regularly in the future.

2.2. OCR analysis

The analysis of OCRs on managerial impact has a long tradition in service management and marketing (Balasubramanian and Mahajan, 2001; Decker and Trusov, 2010; Timoshenko and Hauser, 2019). Table 2 gives an overview of often cited references with pursued tasks from identifying the strengths and weaknesses to understanding which topics are of customer interest. In one of the first analyses in marketing, Decker and Trusov (2010) extracted frequently used words in OCRs for products and related them to (assumed) positive or negative attribute levels. For example, if the product under investigation is a mobile phone, the levels are small or large for the attribute size, high or low for weight, camera quality, voice quality, and battery quality. A negative binomial regression model was used to measure the impact of the occurrence of these words (as proxies for high or low attribute performance) on the overall star rating. The measured impacts reflect the importance of the attribute.

The analyses provided valuable insights for developers and retailers. Qi et al. (2016) and Xiao et al. (2016) modified Decker and Trusov's (2010) approach by applying a combination of conjoint analysis and the Kano model to OCRs in a similar fashion. As an alternative to Decker and Trusov's (2010) approach, Rese et al. (2014) – as discussed earlier – applied a lexicon-based approach based on extended TAM to identify the strengths and weaknesses of the products, services, and technologies investigated.

Hartmann et al. (2023) discussed methodological issues for learning categorizations based on samples of categorized comments. They focused on the question of which methods best predict the overall star ratings of OCRs based on the corresponding comment (as is usual in sentiment analysis). Three alternative approaches were compared: lexicon-based sentiment analysis (LIWC, VADER), traditional machine learning based on word or n-gram occurrences in the comment (Naïve Bayes, Random Forest), and Transformer-based transfer learning (pre-trained and then fine-tuned neural networks). They concluded that the estimated category predictions were valid, and that transfer learning was superior in many cases to the other methods based on an analysis of 272 large-size OCR samples. Consequently, in our methodology, we rely on LLMs and Transformer-based transfer models. In the next two

Table 2
Overview of pursued tasks and proposed methodologies for OCR analysis in marketing and service management.

Reference	OCR dataset	Pursued task	Applied methodology
Decker and Trusov (2010)	n = 20,419 cell phone OCRs with pro/con summaries and a star rating	Identify cell phone attributes from pro/con summaries and the impact of their valence on the star ratings	Extraction of attributes and their valence (e.g., size+, size-, look+, look-), negative binomial regression
Rese et al. (2014)	n = 480 online shop OCRs and n = 275 offline shop OCRs (text, rating)	Predict technology acceptance construct scores from text	Categorization of comments based on a lexicon-based approach, validated by traditional surveys
Büschken and Allenby (2016)	n = 696 restaurant OCRs, n = 4467 hotel OCRs	Identify frequently discussed topics in OCRs	Topic modeling via Latent Dirichlet Allocation
Qi et al. (2016)	n = 679,422 cell phone OCRs (text, rating)	Identify cell phone attributes from pro/con summaries and the impact of their valence	Extraction of attributes by experts and their valence by sentiment analysis
Xiao et al. (2016)	n = 2245 cell phone OCRs with pro/con summaries and a star rating	Identify cell phone attributes and the impact of their valence (+, -) on the star ratings, Kano categories	Extraction of attributes and their valence (e.g., size+, size-, look+, look-), ordered choice model analysis
Kübler et al. (2020)	n = 27,956 comments on 48 brands on Facebook, 5 YouGov mindset brand-day metrics	Predict YouGov mindset metrics (e.g., awareness, satisfaction) from Facebook comments by customers	Sentiment analysis of comments, estimation of vector autoregressive models (VAR) for each metric
Yang et al. (2019)	n = 12,000 customer posts on Facebook for 41 companies, categorized by MTurk workers	Analyze user-generated content (UGC) on Facebook Business Pages: Categories and frequencies	Categorization of sample posts by 5 MTurk workers: 7 categories (e.g., product/service quality complaint); train/test a model
Timoshenko and Hauser (2019)	n = 115,099 oral-care OCRs (text, rating)	Identity relevant customer needs by informative OCRs	Categorization of sample OCRs: “informative” vs. “non-informative”; train/test a model
Zhang et al. (2021)	n = 10,000 electronics, beauty, home, kitchen OCRs (text, rating)	Identify OCRs with innovative improvement ideas	Categorization of sample sentences in OCRs: “not very …” to “very innovative”, train/test a model
Kim et al. (2022)	16 datasets with product category OCRs (text, rating)	Compare topics with high- and low-rated OCRs for each product category	Word frequency analysis with high- and low-rated OCRs for each category
Lee et al. (2022)	n = 4,783,669 metaverse OCRs	Predict sentiment scores for metaverse apps from OCRs	Sentiment analysis (VADER)
Bouschery et al. (2023)	n = 20 air pump OCRs (text, rating)	Identify attributes that impact the star ratings positively and negatively	Summary text generation by an LLM (GPT-3)
Hartmann et al. (2023)	272 publicly available datasets with product OCRs	Predict sentiment scores from text	Sentiment analysis (VADER, BERT)
Kumari et al. (2024)	n = 17,136 Metaverse OCRs	Identify attributes with a high impact on consumer intention to use Metaverse platform services	Topic modeling of OCRs (Latent Dirichlet Allocation) to derive topics/attributes followed by regression (usage intention)
Praveen et al. (2024)	n = 1,031,478 hotel OCRs, a subsample rated by students as pos. or negative	Predict sentiment of OCRs (positive/negative), identify frequently discussed topics	Sentiment prediction and topic modeling (LLMs like BERT topic, Falcon-7B topic, GPT-2)
Zhang and Xu (2024)	n = 96,322 hotel reviews (text, rating)	Identify hotel service attributes with a high impact on customer satisfaction to improve services	Topic modeling of OCRs (Latent Dirichlet Allocation) to derive topics/service attributes followed by regression
This paper	8 datasets with online shop, social network, and streaming app OCRs	Predict technology acceptance construct scores from OCRs	Categorization of comments with respect to TAM constructs by experts, train/test an adapted multilingual BERT model, validated by traditional surveys

sections, we briefly introduce the underlying modeling approach and our adaptation to measure technology acceptance from OCRs.

3. Background: Text categorization using LLMs and transfer learning

3.1. LLMs and the transformer architecture

Since the introduction of the Transformer architecture in machine learning in 2017 (Vaswani et al., 2017), LLMs and transfer learning have become the gold standard for language understanding and generation (Chang et al., 2024). LLMs are general-purpose, artificial neural networks that are based on the well-known encoder–decoder principle in machine learning (Ackley et al., 1985; Rumelhart et al., 1987). Fig. 2 provides a visual representation of this principle: the network consists of successive layers of nodes that are interconnected. The encoder element of the network transforms high-dimensional data from input nodes to low-dimensional representations in the bottleneck layer nodes. The decoder element generates high-dimensional data in the output nodes from these low-dimensional representations. For network training, the transformation parameters are iteratively improved from a random starting solution by large samples of given input and output data pairs,

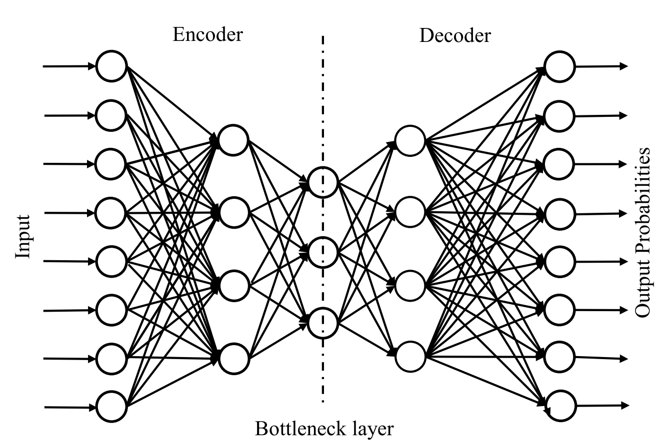

Fig. 2. Encoder–decoder principle in machine learning (adapted from Rumelhart et al., 1987).

so that the calculated outputs from the given input data are as close as possible to the corresponding given output data.

In its simplest form, the encoder–decoder principle is trained with given pairs of identical input and output data to estimate the relation of high-dimensional representations to (unknown) low-dimensional representations without major loss of information (Kramer, 1991). Based on a standardized data matrix with n observations (rows) and m columns (variables), the number of input and output nodes of the network is set to m, and the number of bottleneck nodes is set to a small number that reflects the desired low dimensionality. Then, the network parameters are iteratively trained based on the n observations (the rows) as input and identical output data. Kramer (1991) showed that, in many cases, the achieved results are comparable to nonlinear principal components analysis. The estimated relations (network parameters) between the high-dimensional and the low-dimensional representations can be used to predict meaningful low-dimensional representations (the values in the bottleneck layer nodes), even for (new) inputs.

Over the years, the encoder–decoder principle has established itself as the superior architecture for many machine learning tasks, especially when text (as a sequence of words or word pieces) has to be translated from one language to another (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2015). Moreover, the well-known Transformer architecture (Vaswani et al., 2017) is based on this powerful encoder–decoder principle. The architecture allows us to pre-train LLMs, which can then be fine-tuned – as so-called transfer models – for many natural language processing tasks, such as machine translation, natural language inference, next sentence prediction, paraphrasing, question answering, reading comprehension, sentence completion, sentence acceptability judgment, sentiment analysis, text categorization, and text generation (Raffel et al., 2019).

The left-hand side in Fig. 3 (adapted from Vaswani et al., 2017) illustrates the main components of this Transformer architecture based on the encoder–decoder principle:

- Input and output data are natural language texts (sequences of words or word pieces) that are converted by so-called tokenizers into sequences of tokens as indicators for words or word pieces according to a vocabulary (list of frequent and meaningful words or word pieces in languages).
- The tokens and their positioning (sequence information) are then converted into vector representations in the input and output embedding layers of the network. Here, so-called embedding tables from other models can be used for this coding process.
- Nx successive so-called Transformer layers (e.g., Nx = 6) carry out repeated transformations on these vector representations to extract more and more abstract linguistic information. Each transformation layer consists of a so-called attention layer and a feedforward layer. The attention layers are specific to the Transformer architecture. They allow token or node amplifications to be learned depending on the context, as earlier proposed by Bahdanau et al. (2014). However, in contrast to the former propositions, the Transformer architecture realizes this in a feedforward manner and, therefore, requires less training time.
- An optional so-called un-embedding layer converts the final vector representations back into a probability distribution over the tokens.

To pre-train LLMs with the Transformer architecture, large text corpora were used as a basis for input and output data pairs. Among those used, the 2014 Workshop of Machine Translation (WML) dataset should be mentioned with about 4.5 million English-German sentence pairs (Vaswani et al., 2017), the Toronto BooksCorpus with 800 million, and the English Wikipedia corpus with 2500 million words (Devlin et al., 2018; Raffel et al., 2019). From these text corpora, samples of input and output text pairs were constructed to train a general-purpose LLM. Typical pairs for training reflect tasks such as the restoring of corrupted text (text with masked words as input, text without these masked words as output), a machine translation task (with the same text in two different languages as input or output text pairs, see Raffel et al., 2019).

It should be noted that pre-training an LLM places enormous demands on memory volume and computing time, due to the large number of model parameters and the volume of input and output text pairs needed for training. For example, Vaswani et al.'s (2017) basic LLM ("base") makes use of a word-piece vocabulary with 25,000 tokens and allows up to 512 tokens as input and output text (text with up to 512 word pieces). With Nx = 6 Transformer layers in the encoder and in the decoder elements, this produces 65 million network parameters to be estimated. The largest LLM ("big") with 1024 tokens as input and output is considered to have 213 million network parameters to be trained

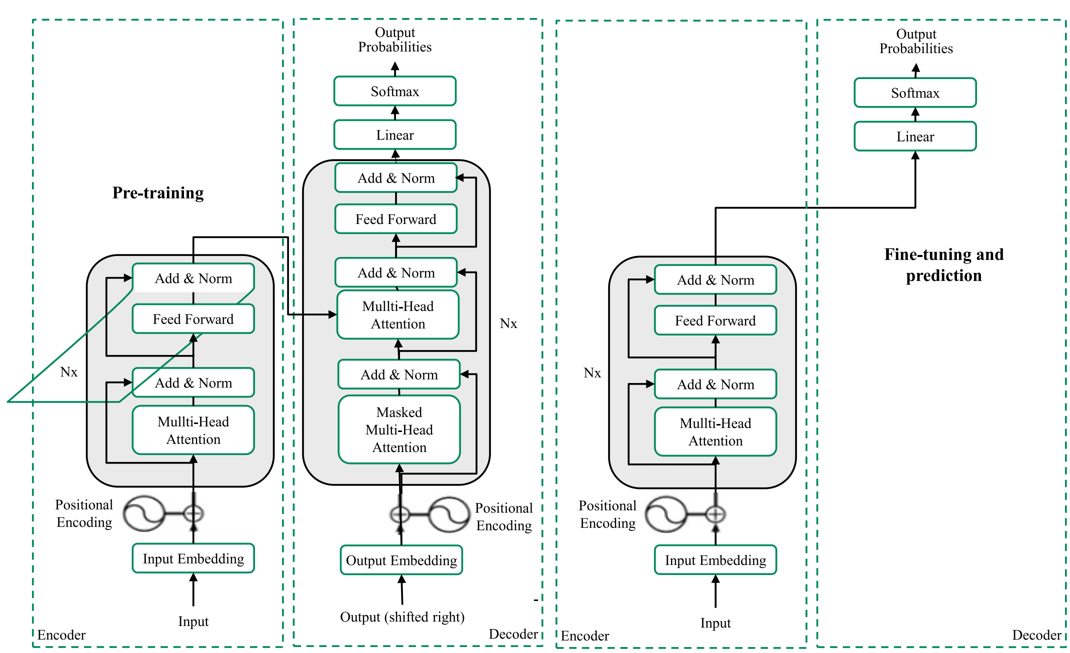

Fig. 3. Illustration of the main components of the Transformer architecture (on the left-hand side, adapted from Vaswani et al., 2017) and its application as a transfer model to text categorization (on the right-hand side).

(Vaswani et al., 2017). Training on a machine with eight Nvidia P100 GPUs took 12 h for the base LLM and 3.5 days for the big LLM. Further developments of LLMs with much more network parameters impose even higher demands on memory volume and computing time.

3.2. Text categorization using transfer models based on the transformer architecture

A major advantage of LLMs is that they can be used for other natural language tasks (e.g., text categorization) than those for which they have been originally trained (e.g., machine translation, language inference, and question answering). For example, the ubiquitous BERT (Bidirectional Encoder Representations from Transformers; see Devlin et al., 2018) is a widespread family of LLMs that mainly consist of the encoder element of the Transformer architecture. BERT was originally trained for language inference with large datasets, but, after fine-tuning to fulfill a new natural language task with much smaller data sets, it is mainly used as a transfer model for text categorization (e.g. for sentiment analysis). The right-hand side in Fig. 3 demonstrates the pre-training and fine-tuning process in the modified Transformer architecture. Input texts are transformed using tokenizers, embedding layers, attention layers, and feedforward layers. A tokenizer automatically prepares the inputs for the model. This typically includes: (1) Splitting words into sub-words (so-called tokens) and assigning them integer IDs with an additional integer representing the position in the sentence; (2) Adding new tokens to the sentence such as "start of the sentence" and "end of the sentence", where the final hidden state is typically used for classification. The embedding layer maps each unique token into a high-dimensional vector representation (for BERT: 768 dimensions, see Devlin et al. 2019). These embeddings transport semantic meaning, such that related words are close in the embedding space.

Attention layers adjust the embeddings on tokens dynamically based on the sentence they occur in. This is important to distinguish words with multiple meanings (e.g., "mouse" can be a device or a mammal), or to account for important relations, such as negations (e.g. "not easy"). Attention layers do this by taking the embedding of all other words that surround the word, scaling it with the learned attention weight matrix, and updating the target embeddings. Feedforward layers add important non-linearity to the model and allow for more complex interactions. They consume the updated embeddings and transform them so they can be used in subsequent attention layers.

They then directly lead by simple transformations to output probabilities for text categories (e.g., star ratings for an OCR). Like many other LLMs, first BERT models were trained by using the Toronto BooksCorpus (containing 800 million words) and the English Wikipedia corpus (with 2.5 billion words). However, unlike encoder-decoder models that train on pairs of input and output texts, BERT uses a bidirectional Transformer architecture. It is trained with two objectives: masked language modeling, where random words in a sequence are masked and predicted, and next sentence prediction, where the model predicts whether two sentences are consecutive in a text. This bidirectional analysis has proven advantageous, especially for tasks like text categorization, where understanding the context of a word or sentence from both directions is crucial (see Devlin et al., 2018 for details). Recently, BERT has been extended to allow sentiment analysis with multilingual texts (especially applicable for texts in Dutch, English, French, German, Italian, and Spanish), and it has demonstrated its superiority over many other LLMs for tasks such as question answering and language inference, without substantial task-specific architecture modifications (Devlin et al., 2018), and for multilingual sentiment analysis (Hartmann et al., 2023; Manias et al., 2023).

General-purpose and specific (for sentiment analysis) BERT LLMs with pre-trained parameters are available for programmers in Google's Keras and Hugging Face's Transformers packages for Python (Chollet, 2021), and they can be further fine-tuned to become transfer models by providing additional task-specific pairs of input data and corresponding categories. We make use of Hugging Face's multilingual BERT for sentiment analysis as a basis for fine-tuning our research methodology. However, in contrast to former OCR analyses, we use BERT not to predict sentiment categories themselves but, for the first time, to predict extended TAM construct scores.

4. Research methodology

4.1. Datasets

In order to demonstrate that transfer models can complement or replace the traditional measurement of extended TAM constructs by customer surveys – our main research endeavor – we scraped larger OCR datasets from Google Play Store. Table 3 documents these datasets (including one dataset, Ikea2, with OCRs collected during a survey in a laboratory setting, as discussed later). Across all datasets, the focus is on OCRs for shopping apps (Ikea, Baur), social media apps (Facebook), and video streaming apps (Disney+, Netflix, Prime). One app (Facebook) received a rather low mean star rating, and one app (Baur) obtained a rather high rating.

It should be mentioned that our datasets contain no personally identifiable customer information: The stored OCRs contain the date and the content of the review, the star rating, and a successive customer number. Our analyses conform with article 89 GDPR with purposes that fall under recital 159 ("processing for scientific research purposes", see gdpr-info.eu/recitals/no-159).

The first dataset, Ikea1, shows – similar to Disney+, Netflix, and Prime – a moderate mean star rating with the usual U-shape in frequencies: many 1-star and 5-star ratings, but few 2-, 3-, and 4-star ratings. The reviewed Ikea app was provided by Ikea (www.ikea.com), a Swedish multinational conglomerate that designs, manufactures, and sells ready-to-assemble furniture and various other goods. Besides Google Play Store (Android), this app can be downloaded from Apple's App Store (iOS) and offers customers – according to the company's website – the opportunity to not only buy furniture at home or to check stock availability but also to create shareable shopping lists and to scan products in-store to learn more about them. The app has received over ten million downloads from Google Play Store and was rated more than 159,000 times by its users with an average rating of 4.5 stars (on the usual 1-star to 5-star overall OCR rating). However, for our OCR analysis, we focus on all available OCRs with an additional natural language text (n = 5356) across three years (from 11/2019 to 12/2022). It should be noted that the OCRs with a comment showed, on average, a much lower rating compared to all OCRs.

For the 5356 OCRs in the dataset, we collected human expert categorizations with respect to the six TAM constructs. The three experts used for this purpose were advanced master-level students in computer science and marketing, knowledgeable in TAM/UTAUT and related technology acceptance models. They were given sample questionnaires (Rese et al., 2014, 2017; Schreiber, 2020) with typical indicators for the six extended TAM constructs used in many surveys (see Table 1). We asked them to categorize each OCR comment accordingly, using 1-star to 5-star ratings (comparable to a 5-point Likert scale). The 1-star and 5-star ratings should only be used if a clear negative or positive indication of the construct is in the comment. The OCR's overall star ratings were disguised during this categorization task. However, due to the often clear expression in the comments by words such as "super", "marvelous", "horrible", or "de-installed immediately", "would never use again", "looking forward to next use", we expected the star ratings on two TAM constructs – attitude toward using and the behavioral intention to use – to have similar categorizations as the (not shown) overall star ratings with many 1-star and 5-star ratings. Table 4 summarizes these (to some extent expected) categorizations of the two constructs by the three experts. Notably, the one TAM construct – perceived informativeness – received few 1-star and 5-star ratings from the human experts since the OCR comments referred to this construct

Table 3
Datasets with OCRs used for predictions of the extended TAM scores (downloads according to play.google.com).

Dataset	Ikea1	Ikea2	Ikea3	Baur	Disney+	Facebook	Netflix	Prime
Logo								
OCRs	n = 5356	n = 275	n = 480	n = 3091	n = 27,160	n = 29,053	n = 64,233	n = 37,373
Frequ-ency of overall star ratings	1: 1,620, 2: 417, 3: 281, 4: 437, 5: 2601	1: 7, 2: 39, 3: 94, 4: 113, 5: 22	1: 229, 2: 60, 3: 67, 4: 51, 5: 73	1: 367, 2: 124, 3: 118, 4: 282, 5: 2200	1: 9,534, 2: 2,614, 3: 2,372, 4: 2,670, 5: 9970	1: 20,607, 2: 1,662, 3: 1,288, 4: 1,151, 5: 4345	1: 14,596, 2: 4,875, 3: 6,153, 4: 8,154, 5: 30,455	1: 7,948, 2: 3,093, 3: 3,247, 4: 5,164, 5: 17,921
Mean	∅ = 3.370	∅ = 3.378	∅ = 2.329	∅ = 4.237	∅ = 3.034	∅ = 1.863	∅ = 3.545	∅ = 3.589
OCR time span	11/2019 to 12/2022	06/2013 (part of a survey)	11/2011 to 11/2013	11/2016 to 03/2023	10/2019 to 03/2023	05/2021 to 03/2023	06/2012 to 03/2023	11/2016 to 03/2023
Down-loads	>10 Mio.	>10 Mio.	>10 Mio.	>1 Mio.	>100 Mio.	>5 Bill.	>1 Bill.	>500 Mio.

Table 4
Frequencies of star ratings contained in the n = 5336 OCRs ("overall", listed in the first row) and of the expert star ratings (n = 5336 OCR comments rated by n = 3 human experts) for the extended TAM constructs (listed in the rows below "overall") as well as mean star rating, correlation with the overall star rating, and inter-rater reliability.

	Frequencies of star ratings					Mean star rating	Overall correlation	Cron-bach's α	Krippen-dorff's α
	1-star	2-star	3-star	4-star	5-star				
Overall (disguised to experts)	1602	417	281	437	2601	3.370	1	–	–
Perceived informativeness	104	42	15,718	54	150	3.006	0.1308	0.858	0.770
Perceived enjoyment	814	157	8893	819	5385	3.610	0.6092	0.911	0.850
Perceived usefulness	5416	606	3787	2774	3485	2.895	0.7972	0.962	0.930
Perceived ease of use	587	227	11,284	1902	2068	3.289	0.3880	0.793	0.670
Attitude toward using	5413	971	620	965	8099	3.334	0.9686	0.991	0.982
Behavioral intention to use	3295	1047	2207	2927	6602	3.530	0.8544	0.994	0.989

rarely and the experts followed our instruction to use the 3-star rating in cases where the comment didn't refer to the construct's meaning. So, e.g., comments containing n-grams like "super easy to use", "works perfectly", "helps a lot", "I like it" led to 4- or 5-star expert ratings for perceived enjoyment, perceived ease of use, perceived usefulness, attitude toward using, or behavioral intention to use, but only to a 3-star rating for perceived informativeness.

It should be mentioned that two constructs, perceived informativeness (98.2% 3-star ratings) and perceived ease of use (70.5%), received a high percentage of evaluations with the neutral category. Due to these two constructs being not discussed in many OCRs, the experts followed our instructions and evaluated them in these cases as "don't know/neutral". However, in OCRs, such unbalanced distributions of categories (e.g., U-shared or inverse U-shaped star ratings) are quite common. So, e.g., in Table 3 all datasets show more or less the typical U-shaped distribution for overall star ratings (with many 1-star and 5-star ratings and few 2-, 3-, and 4-star ratings) as also noted for other OCR datasets by, e.g., Hu et al. (2006), Asghar (2016), or Kovács (2024). Asghar (2016) and Kovacs (2024) propose to analyze such unbalanced distributions by multi-class approaches (predicting the probability for each rating category separately), as we will do later, before aggregating the five categories by weighted averaging. However, one should keep in mind that the 3-star prediction represents a neutral position.

The quality of the three expert ratings was evaluated by two criteria for inter-rater reliability (see Table 4). One criterion is Cronbach's α (Cronbach, 1951) with a threshold value of 0.7, which is exceeded in all cases. The relative evaluation but not the absolute evaluation is taken into account. Here, the three experts rated very consistently. Hayes and Krippendorff (2007) proposed Krippendorff's α (Krippendorff, 1970) as a second criterion, again, most values were excellent at 0.8 and higher, one was good (≥0.75), and one was at least acceptable (≥0.667). The differences between the reliability measures are not surprising since Krippendorff's α reflects the congruence of absolute, but not relative, ratings, as is the case with Cronbach's α. Therefore, a different rating overall reduces the value in absolute terms. A closer look at the single expert's judgments confirmed that one expert was more critical during the tasks than the other two – in particular, when perceived ease of use and perceived informativeness had to be evaluated. Since the relative evaluation and the correlations in particular are of major importance in our study, we do not consider these low evaluations of inter-rater reliability as problematic.

Since the three experts were students, we wanted to check whether their star ratings were biased by their academic education. We searched for additional raters to validate the students' evaluations from a broader perspective. For this purpose, recently, the usage of silicon samples produced by generative artificial intelligence chatbots was proposed since they allow to mimic answers of typical customers as well as of human experts in many areas (see the overview in Arora et al., 2024 or Sarstedt et al., 2024). So, e.g., Arora et al. (2024) applied OpenAI API for GPT-4 (OpenAI, 2023) as a chatbot to simulate answers of a synthetic sample of n = 605 respondents in a conjoint study and compared the derived results with the results based on answers of a human sample of n = 605 respondents. The synthetic sample was generated according to the characteristics of the human sample (defined by gender, age, income, urbanicity, education, and ethnicity, used as prompts for GPT-4 during the answer generation process) and showed promising results: Concerning derived attitudes and purchase likelihood, the derived results of the synthetic sample "mimic the direction and the valence" of the results of the human sample (Arora et al., 2024, p. 36). Guo et al. (2023) invited 17 volunteers to evaluate ChatGPT answers to questions (and related tasks) in comparison to human expert answers in areas like finance or psychology. The pairs were derived partly from publicly available question-answer datasets and partly constructed by the authors using definitions and applications of concepts from wiki sources like Wikipedia and BaiduBaike. The ChatGPT answers to these questions were generated by manually inputting the questions (or tasks) into ChatGPT's input box and storing ChatGPT's answer together with the human expert answer. The volunteers were iteratively shown questions (or tasks) together with the two stored answers and had to decide which answer was more helpful (without knowing which of the two the chatbot's answer was). Surprisingly, the ChatGPT answers were generally considered to be more helpful, especially in areas like finance and psychology (Guo et al., 2023). Moreover, an additional evaluation by real experts in these areas demonstrated that – indeed – the ChatGPT answers were more concrete and specific. Guo et al. (2023) concluded that – at least in areas like finance and psychology – ChatGPT is able to replace an expert in a question-answer setting.

Based on these findings, we used the chatbot OpenAI ChatGPT 4o (OpenAI, 2024) to generate non-human expert ratings of the six constructs based on the comments in dataset Ikea1: ChatGPT received the operationalization of the six TAM constructs according to Table 1 as context and the OCR comments as data. Then, ChatGPT was asked to evaluate all six constructs on a 5-point Likert scale based on the shown information. Appendix A of this paper shows a sample prompt for this task, and Appendix B sample answers generated by ChatGPT. The task for the chatbot was close to the task for the three human experts as discussed above. However, due to the context remembering restrictions of ChatGPT, the construct definition context had to be inputted repeatedly (before each evaluation of a comment), which led to a long interaction and computation time of about 14 h for the 1000 most recent comments in the dataset Ikea1. The resulting ChatGPT construct evaluations for these 1000 comments were compared to the corresponding evaluation of the three human experts, leading to the following Cronbach's α values (with the four ratings as items): 0.538 for perceived informativeness, 0.853 for perceived enjoyment, 0.845 for perceived usefulness, 0.633 for perceived ease of use, 0.954 for attitude toward using, and 0.957 for behavioral intention to use. The rather low Cronbach's α values for two constructs (perceived informativeness and perceived ease of use) in both comparisons could be traced back to the problem that ChatGPT sometimes evaluated these constructs positively or negatively even when the comment contained no information on these constructs. For example, comments containing words like "super" or "extraordinary" not only led to high ratings for perceived enjoyment, perceived usefulness, attitude towards using, and behavioral intention to use, but also led to high ratings for these two constructs. We modified the prompt and added the instruction "Use the '3 (don't know, neutral)' answer if the comment did not reflect the items of a construct", but this modification did not solve the problem. Here, further research is needed concerning this obvious Halo effect in ChatGPT's answers. Nevertheless, the high reliability values for all constructs demonstrate that the expert ratings by the three human experts and by ChatGPT were quite similar.

Additionally, for receiving ratings for all OCRs, we used a Python implementation for accessing the chatbots ChatGPT-4o (as above) and Gemma2-9B (Google's generative artificial intelligence chatbot) in a zero-shot classification setup (with similar prompts as in Appendix A, see the Online Appendix for more details). The outputs from LLMs are often non-deterministic, i.e. the same input produces two different outputs. For the sake of text classification our approach does the following to produce more deterministic results: 1. We set the temperature parameter to 0.0, which reduces the randomness by selecting the most probable token; 2. We define a so-called structured output model that defines how the LLM should structure its answer. This model is defined using Pydantic, a Python library for data validation and settings management, and defines the output as a single label. We subsequently define a base prompt for the LLM containing a description of the construct, the possible output values and their meaning, an instruction to rate an OCR accordingly, and a placeholder for the OCR. We then use a function that receives OCRs; for each OCR the function replaces the placeholder in the base prompt with the OCR, sets the temperature and structured output, and finally passes the complete prompt to the LLM. For every OCR we then receive a single label. This process is repeated for each construct.

By this, we produced fourth (by ChatGPT-4o) and fifth (by Gemma2-9B) ratings of the six constructs for all 5356 reviews. The three human expert evaluations and the two chatbot evaluations resulted in the following Cronbach's α values (that compare now five ratings for the 5356 reviews as items): 0.622 for perceived informativeness, 0.870 for perceived enjoyment, 0.958 for perceived usefulness, 0.799 for perceived ease of use, 0.983 for attitude toward using, and 0.965 for behavioral intention to use. Again, the high reliability values demonstrate that the expert ratings by the three students and the two additional chatbots were quite similar.

To extend our insights on Ikea and its apps, two additional datasets were made available for our analysis (see Table 4). Both datasets refer to the so-called Ikea Place app, a former app implementation provided by Ikea. The app allowed users to virtually place true-to-scale 3D models in their very own space using augmented reality technology. This app was promoted, before its 2017 relaunch, under the name mobile Ikea catalogue app with augmented reality features. Acceptance analyses and OCR analyses for this app were published in Rese et al. (2014, 2017), and Schreiber (2020). The corresponding datasets were made available by the authors on request. The dataset Ikea2 summarizes a laboratory data collection that has been conducted at a German university with n = 275 participants. Most participants were German undergraduate university students. The gender distribution in the sample was 58.9% male and 41.1% female. On average, the participants were 22.1 years old (18–34: 98.9%). Rese et al. (2014) state that the gender distribution (and to a lesser extent the age distribution) matched quite well the characteristics of smartphone users that scanned QR codes at that time, a target segment that Ikea expected to reach with this app (personal communication with Ikea managers in Berlin early in 2013 when the survey was designed). Experience with the app was not required. During data collection, the participants were briefly introduced to the functionality and use of the app. They became acquainted with the additional virtual content by being asked to search for specific furniture using the Ikea printed catalogue and the app. After this interaction, the participants were required to answer an extended TAM questionnaire comparable to Table 1. The questionnaire additionally contained an open question at the beginning to receive a star rating and a comment, similar to an OCR rating.

The dataset Ikea3, on the other hand, is a collection of OCRs of real users of the Ikea app collected (scraped) from Google Play Store at about the same time as the laboratory data collection took place. At the time in question, the app was not widely used, so only n = 480 OCRs were available that, besides the overall star rating, also contained a comment.

4.2. Three-step approach to train, test, and validate the transfer models

Based on our background and the datasets discussed, our research methodology to answer the research question from Section 1 ("*Can transfer models complement or replace traditional measurement of extended TAM constructs with customer surveys?")* can be described as follows:

In a *first step*, we fine-tune a pre-trained LLM for 5-star sentiment prediction (multilingual BERT for sentiment analysis) to form six transfer models to predict scores from comments, one model for each of the six extended TAM constructs. For this purpose, the Ikea1 dataset is used with n = 5356 OCRs, construct-specifically categorized by the three human experts on a 1-star to 5-star rating (see Section 4.1). Take for example the following OCR natural language text for which we want to determine the rating of the construct perceived informativeness:

Text = "*How can I store my Ikea Family membership? The app only offers the option of becoming a new member. I think the amount of information that the app and website now offer is below average. Before the update, you could still see the weight of duvets (per square meter), but not anymore. Information about promotions in my furniture store cannot be displayed either. Transparency is something else. The focus here is more on nice pictures*" (translated from German to English, the original German text can be found in Appendix C).

Based on an, here, identical selection of the 1-star category by the three human experts, the machine learning model should learn to classify perceived informativeness into the 1-star category (low informativeness). This process is repeated for every review and construct. We use a standard 80:20 split for training and testing. We determine the split once randomly for all models, so they are all evaluated using the same texts. We follow the recommended fine-tuning approach by Devlin et al. (2019) and on Hugging Face's website (https://huggingface.co/docs/transformers/en/training):

1. Reinitialize a new dense output layer (here: 5 units for the 1-star to 5-star categories)
2. Fine-tune the complete model
3. Use a very low learning rate (here: 0.00001)
4. Train for only few epochs (here: 3)
5. Use the recommended loss function for the task at hand (here: cross-entropy)

Additionally, we did not change the hidden activation functions (by default Gaussian Error Linear Unit), as this would require complete retraining, nor did we define an activation function for the output layer (by default this is a linear activation function) since we require a logit output. The logit output is then scaled to probabilities using a softmax function. If we define the right side of Fig. 3 as BERT_PI (the fine-tuned model for perceived informativeness) and pass it the example text, we would receive the following output probabilities for the five categories: BERT_PI(text) = [[0.9864268, 0.00440165, 0.00654561, 0.00136861, 0.00125729]].

We compare the train and test accuracy of the transfer models with the accuracy achieved by traditional machine learning (Naïve Bayes) and the lexicon-based approach (Rese et al., 2014). Naïve Bayes is a simple, popular, and well-known standard approach for supervised text classification (see, e.g., Dey et al., 2016): Texts are described by binary indicators (indicating, e.g., whether a text contains pre-specified words or n-grams). From the train sample, the model learns how probable words or n-grams are in each category. Relying on Bayes theorem for conditional probabilities, these probabilities are used to predict how probable categories are when a text contains specific words or n-grams (Dey et al., 2016). The term "naïve" reflects the underlying assumption that the occurrence of words or n-grams in a text is independent from the occurrence of other words or n-grams (which is problematic but allows calculation of joint probabilities by multiplication, Dey et al., 2016). The train and test samples are used to measure the accuracy of the model which later can be used to predict categories also for up-to-now uncategorized texts. Due to its simple functionality and demonstrated high accuracy in many text classification cases, Naïve Bayes is often used as a benchmark for comparisons with new approaches (here: the transfer models). Here, it should be mentioned that we compared our transfer models also with other text classification procedures but omitted this discussion here due to space restrictions (results are available from the authors at request, the other procedures provided inferior results).

Note that the transfer models and the Naïve Bayes approaches predict probabilities for the five response categories for each construct (the 1-star to 5-star ratings). These predictions are transformed into predicted real-valued star ratings (ranging from 1.0 to 5.0) in the usual manner, using weighted averages (the sum of the stars times the corresponding probabilities divided by the sum of the probabilities). More details can be found in the corresponding online appendix, where we provide the codes used to fine-tune the six models (see, e.g., the file bert_base_multilingual_ATT.py for training the transfer model for the construct attitude towards using).

In the *second step*, we validate the six transfer models using the three Ikea datasets (Ikea1, Ikea2, and Ikea3, see Table 3) by applying structural equation modeling according to Fig. 1 (a) to observed ratings of the six constructs (Ikea1 rated and Ikea 2 rated) as well as to (b) to predicted construct scores based on the six transfer models (Ikea1 predicted, Ikea2 predicted, and Ikea3 predicted), checking for construct and discriminant validity whenever possible (Ikea1 rated and Ikea2 rated) and comparing the estimated coefficients and fit criteria across the five estimations. Ikea1 rated is based on the construct-specific star ratings by the human experts (with three ratings as items per construct), Ikea2 rated is based on the construct-specific multi-item responses from the Ikea2 TAM survey (with four to five items per construct according to Table 1). Furthermore, the predicted Ikea1 construct scores are used to discuss the evolution of acceptance over time.

In the *third step*, the six trained, tested, and validated transfer models are applied to the other five OCR datasets (Baur, Disney+, Facebook, Netflix, and Prime) to derive predicted construct scores. Again, structural equation modeling is applied to these scores, and the coefficients and fit criteria as well as the evolution of acceptance over time are discussed.

5. Research results based on the Ikea datasets

5.1. Training and testing the six construct-specific transfer models

In the first step of our methodology, we developed six transfer models based on the Ikea1 dataset with n = 5356 comments and their construct-specific star ratings by the three human experts (resulting in n = 16,068 observations per construct). For this purpose, we applied Python and multilingual BERT for sentiment analysis, made available through Hugging Face's Transformers package (Devlin et al., 2018; Wolf et al., 2020). To train the six transfer models, a training subsample of the Ikea1 dataset based on 80% of the OCR comments and their corresponding construct-specific categorizations (ratings on a 1-star to 5-star scale) were used. The other subsample (based on 20% of the OCR comments) was employed for testing. The predicted probabilities for each category (the 1- to 5-star ratings) were used as weights to predict real-valued star ratings in the [1.0, 5.0] interval via weighted averaging. Fine-tuning and prediction activities were rather slow, with standard laptop computation times per transfer model of approximately 9.1 min for training and approximately 1.6 min for testing. Table 5 shows the accuracy of training and testing these six transfer models in the column transfer model.

Overall, the correlations between the observed and predicted extended TAM scores are rather high. Their accuracy is especially high when we compare them with two alternative approaches: (1) prediction according to the lexicon-based approach proposed by Rese et al. (2014), and (2) traditional machine learning prediction based on Naïve Bayes (see Section 4.2). For these two approaches, in contrast to the transfer modeling approach, common pre-processing of the comments is needed. We removed digits, punctuation symbols, and accent marks, converted everything to lower cases, and applied word stemming (see Rese et al., 2014 for details). Construct score predictions according to the lexicon-based approach were calculated by summing up and normalizing partial scores for pre-specified stemmed word n-grams as indicators of lower or higher construct scores. For these calculations, the dictionary with partial scores developed by Rese et al. (2014) for the six construct scores was used (with partial scores for, e.g., the construct perceived of use: +1 for "easy", -1 for "not easy", +1 for complicated, -1 for "not complicated" and so on). Due to the available dictionary and its partial scores a split of the data into train and test is not needed (for further details see the discussion in Rese et al., 2014). However, for Naïve Bayes, as for the transfer model approach, a split of the Ikea1 dataset into 80% for training and 20% for testing is applied. The model for each construct was trained and tested separately. The standard laptop computation time for training was approximately 4.5 min per construct and approximately 3.7 min for testing. Comparable to the transfer models, the predicted softmax probabilities for the five target categories in each model (the 1-star to 5-star ratings) were used as weights to predict a real-valued star rating as a weighted average. Table 5 shows the accuracy of these two approaches and demonstrates

Table 5
Accuracy of predictions for Ikea1, measured as correlations between observed and predicted extended TAM construct scores (n = 5356 OCRs rated by 3 experts, resulting in n = 16,068 observations per construct).

Extended TAM construct	Lexicon-based approach (Rese al. 2014, 2017)	Naïve Bayes		Transfer Model	
		Train (80%)	Test (20%)	Train (80%)	Test (20%)
Perceived informativeness	0.0735	0.5406	0.4603	0.7345	0.7399
Perceived enjoyment	0.2404	0.5094	0.4773	0.8899	0.8072
Perceived usefulness	0.3371	0.6320	0.5746	0.9390	0.8946
Perceived ease of use	0.2924	0.4662	0.4160	0.6928	0.6077
Attitude toward using	0.5568	0.7690	0.7488	0.9847	0.9722
Behavioral intention to use	0.4851	0.6986	0.6619	0.9706	0.9427

the superiority of the six transfer models over other approaches. Consequently, only the transfer models were further used to predict extended TAM construct scores from OCR comments.

5.2. Validation of the construct-specific transfer models

Based on the observed ratings of the extended TAM constructs by the three human experts in Ikea1 (Ikea1 rated), the observed responses of the n = 275 customers to the TAM questionnaire in Ikea2 (Ikea2 rated), and the predicted scores of the extended TAM constructs by the six transfer models based on datasets Ikea1, Ikea2, and Ikea3 (Ikea1 predicted, Ikea2 predicted, Ikea3 predicted), it is possible to estimate coefficients of the extended TAM from Fig. 1 using SmartPLS and to detect differences among coefficients and fit criteria across the five estimations and to compare them (Ikea1 rated, Ikea1 predicted, Ikea2 rated, Ikea2 predicted, Ikea3 predicted). Fig. 4 summarizes these results with coefficients and R^2 values (variance explained). Regardless of the measurement approach used, the most hypothesized relations from Fig. 1 between the constructs are confirmed. There is a strong relation between attitude toward using and behavioral intention to use as well as between perceived usefulness and attitude toward using. Moreover, perceived ease of use has a positive effect on attitude toward using, and perceived enjoyment has a significant effect on perceived usefulness.

Major differences, however, can be observed between the estimations based on the Ikea1 dataset on one side and the Ikea2 and Ikea3 datasets on the other side concerning the direct effect of perceived usefulness on behavioral intention to use: For estimations based on the Ikea1 dataset, this direct effect is not significant. In addition, the effect of perceived informativeness on usefulness is significant but low. Additionally, we had a closer look at the construct reliability and discriminant validity of the measurements whenever possible (measurements based on Ikea1 rated and Ikea2 rated with multi-item scales due to the expert ratings per construct or due to four to five items per construct in the questionnaire; measurements based on predicted scores only have single-item constructs). Tables 6 and 7 reflect the commonly used quality criteria for this purpose.

In both cases, Cronbach's α (with values larger than 0.7) and high values for the average variance extracted indicate that the constructs are internally consistent. The 3-star ratings for each construct by the expert raters (in the case of Ikea1) across the n = 5356 OCRs and the Likert scale ratings for the four to five indicators of each construct (in the case of Ikea2) across the n = 275 respondents demonstrate correlating values. On the other hand, the low values for the heterotrait-monotrait ratio (<0.85 is proposed, and only the value 0.874 for the pair, attitude toward using and perceived usefulness, slightly exceeds this criterion) indicate that the extended TAM constructs in both estimations (based on Ikea1 rated and Ikea2 rated) demonstrate sufficient difference to be addressed as well defined. Moreover, other quality criteria have been successfully checked (Hair et al., 2017; Ringle et al., 2012).

Moreover, using the timestamps of each OCR, the extended TAM constructs can be developed over time. Fig. 5 visualizes this presentation for the Ikea1 dataset: We averaged the predicted scores of each construct monthly. Additionally, in Fig. 1, two important relaunches of the Ikea app during our observation period are indicated: In February 2021, the app received a complete relaunch with a new look & feel, new login procedures, and a modified order to purchase process. In December 2021, the new shop & go functionality was introduced (the customer scans QR codes when putting articles in the shopping basket in the store, this simplifies the checkout in the store, see https://www.ikea.com/de/de/customer-service/mobile-apps/). Both relaunches were accompanied by customer problems, and app adaptations followed.

A more detailed look at Fig. 3 offers some interesting insights. The

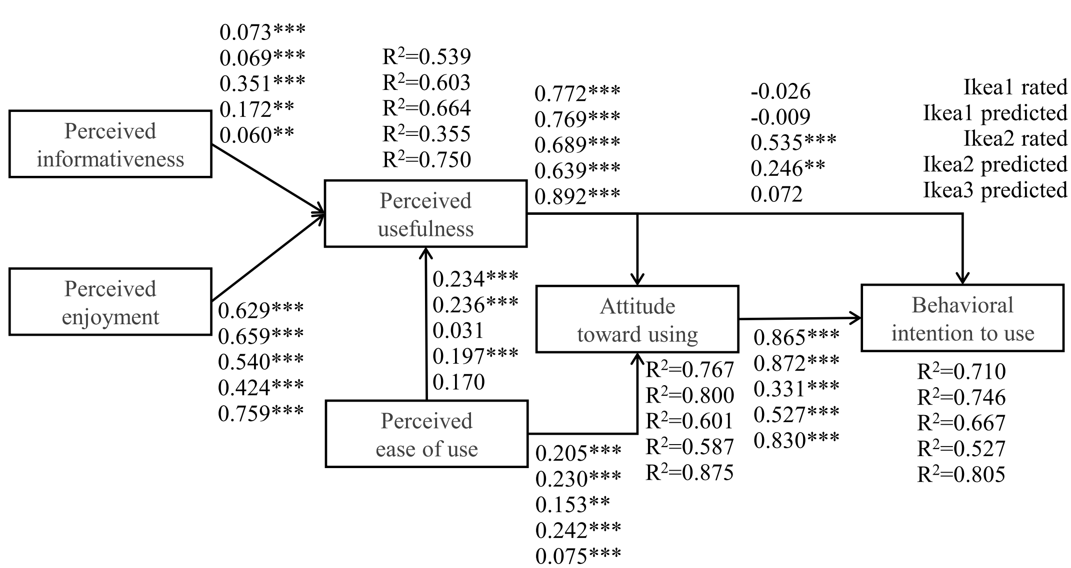

Fig. 4. Extended TAM results for the Ikea datasets across different approaches to measuring the scores (*: p<0.05, **: p<0.01; ***: p<0.001).

Table 6
Construct mean, standard deviation (on a 1- to 5-star scale), reliability, and discriminant validity for the measurement based on Ikea1 rated (n = 5356 OCRs, the three expert ratings per construct are used as items).

	Mean	Stan. deviation	Cron-bach's α	Average variance extracted (AVE)	Heterotrait-monotrait ratio (HTMT)				
					PI	PE	PU	PEOU	ATT
Perceived informativeness (PI)	3.008	0.195	0.858	0.770					
Perceived enjoyment (PE)	3.609	0.993	0.911	0.850	0.105				
Perceived usefulness (PU)	2.906	1.459	0.962	0.930	0.156	0.739			
Perceived ease of use (PEOU)	3.273	0.573	0.793	0.670	0.067	0.227	0.377		
Attitude toward using (ATU)	3.335	1.803	0.991	0.982	0.160	0.706	0.874	0.495	
Behavioral intention to use (BI)	3.525	1.500	0.994	0.989	0.141	0.611	0.728	0.377	0.849

Table 7
Construct mean, standard deviation (on a 1- to 5-star scale), reliability, and discriminant validity for the measurement based on Ikea2 rated (n = 275 participants in the survey, four to five items per construct).

	Mean	Stan. deviation	Cron-bach's α	Average variance extracted (AVE)	Heterotrait-monotrait ratio (HTMT)				
					PI	PE	PU	PEOU	ATT
Perceived informativeness (PI)	2.965	0.447	0.870	0.659					
Perceived enjoyment (PE)	3.667	0.656	0.892	0.757	0.644				
Perceived usefulness (PU)	2.893	1.119	0.917	0.802	0.740	0.834			
Perceived ease of use (PEOU)	3.123	0.654	0.895	0.761	0.447	0.661	0.536		
Attitude toward using (ATU)	3.125	1.106	0.943	0.815	0.687	0.758	0.832	0.558	
Behavioral intention to use (BI)	3.531	0.754	0.914	0.747	0.643	0.758	0.852	0.522	0.805

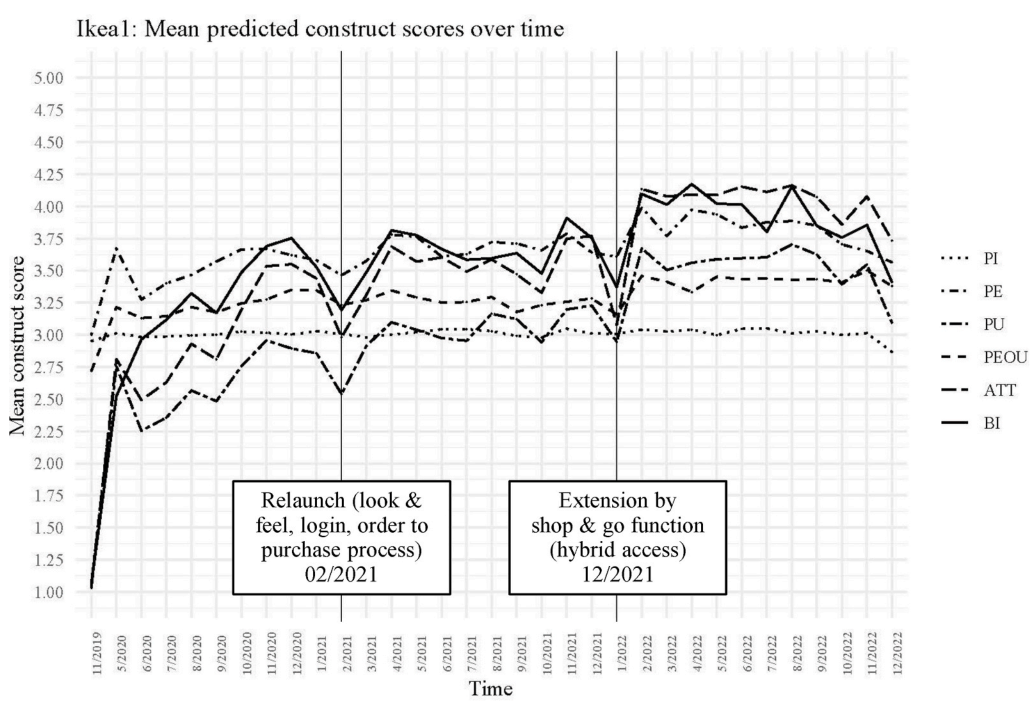

Fig. 5. Predicted extended TAM construct scores over time for Ikea1 (for the abbreviations see Tables 6 and 7).

perceived usefulness of the Ikea app has considerably improved over time whereas the attitude toward using and the behavioral intention to use have stagnated. Since the period largely covers the time of the Corona pandemic, this observation reflects a missed chance for Ikea to gain new customers, at least for the app. Furthermore, perceived ease of use has slightly improved over time but not as much as perceived usefulness. Looking at the categorized OCRs in combination with Fig. 3 shows that even more information is available. Concerning the topics that are discussed in OCRs with low and high perceived usefulness, a comparison cloud (not shown here) revealed that the words and n-grams, such as "browser", "Chrome", "data protection", "registration", and "spying", are very frequent in OCRs with low perceived usefulness. These OCRs reflect the need to use a separate Chrome browser session for registration, which was, for a long time, a major reason why many users did not register and were, therefore, unable to deploy the full functionality of the app (e.g., favorite lists, recently visited articles). Other additional insights regarding the development of the extended TAM constructs over time refer to the highly controversial discussion of the new app design (compared to its predecessor) in the light of low versus high perceived enjoyment and low versus high usefulness due to missing functions.

6. Research results based on the other datasets

In the last step of our methodology, we apply our transfer models to n = 160,910 OCRs concerning five additional platforms/services (see the five non-Ikea columns in Table 5). OCRs were scraped from Google Play Store (Android). The apps range from online shops (Baur) and social networks (Facebook) to streaming services (Disney+, Netflix, Prime). They represent a wide range of popular apps as can be seen from the number of app downloads. Fig. 6 summarizes the predictions of TAM scores and their analysis using SmartPLS. Regardless of the datasets, the hypothesized relations between the constructs are confirmed.

However, once again, the variances explained in the extended TAM are high. Moreover, based on the score predictions, the evolution of the construct scores can be derived from the results. For example, the development of the scores over time for the Baur app, a major German online shop for fashion and home accessories, is presented in Fig. 7. Again, major relaunches of the app are indicated.

Compared to the Ikea app, the extended TAM constructs develop differently over time. Overall, the constructs have low scores in the early years, but also in recent years. A comparison cloud analysis for these two periods shows that, in the early years, the app had some problems due to low functionality and many technical issues, whereas in recent years – with many new users and sales increases due to the Corona pandemic – some new users tried the app and were disappointed. However, perceived enjoyment was often very high and could be related to some design features discussed in the OCRs.

7. Discussion of results and methodological limitations

7.1. Discussion of results

The analysis of the Ikea and other datasets with large OCR samples of apps has demonstrated that valid TAM construct score predictions, solely based on OCRs, are possible. An earlier attempt with a (simple) lexicon-based approach to predict TAM construct scores (Rese et al., 2014) revealed shortcomings concerning the size and significance of the estimated relations between the TAM constructs, the explained variance in the TAM predictions, and the convergent and discriminant validity. In this paper, due to training and testing Transformer-based transfer models, the construct score predictions mostly led to high explained variances, significant relationships between the constructs, and fulfilled quality criteria concerning convergent and discriminant validity. Moreover, beyond this internal validity assessment, comparisons of the construct predictions with (1) human expert as well as generative artificial intelligence chatbot ratings and (2) estimates from traditional TAM measurement (using multi-item measurement of the constructs and sampling of respondents) demonstrated high external validity of the construct score predictions.

It should be reiterated that our modification of a BERT model (Devlin et al., 2018) for sentiment analysis of multilingual comments (Hartmann et al., 2023; Manias et al., 2023), based on Hugging Face's Transformers package for Python (Chollet, 2021) and its fine-tuning based on the human experts' construct evaluations, was responsible for this superiority. It greatly outperformed the lexicon-based approach of Rese et al. (2014) and the Naïve Bayes train/test approach.

In addition, for the first time to the best of our knowledge, it was possible in our analysis to discuss the evolution of TAM construct scores over time. This was due to the modified basis for measurement of using OCRs with timestamps instead of survey responses at a certain point in time. Figs. 5 and 7, based on the OCRs and the version development of the Ikea and Baur apps, demonstrate the usefulness and the validity of this approach. This new possibility allows managers to monitor the quality of their app over time and to control the effects of improvements (e.g., new versions of products, services, or technologies).

7.2. Discussion of methodological limitations

However, our proposed methodology has some important limitations. For example, it is well known that customers writing OCR comments differ from average customers who rate apps but do not write OCR comments (Han and Mikhailova, 2024), and from the average customer (Dixit et al., 2019). We observed this also in our investigation: So, e.g., our Ikea1 dataset showed an average star rating of 3.37 across OCRs with a comment, being much lower than the average star rating of 4.5 across all OCRs (with or without a comment). Dixit et al. (2019) characterize the OCR comment writer as follows: customers for whom OCR comment writing is an important part of their self-concept (ego involvement), customers who want to take vengeance, customers who are used to writing OCR comments (perceived behavioral control), and customers whose friends or family also write OCR comments (subjective norm). When taking OCRs as a basis for acceptance measurement, it should be clear that customers with these motivations are over-represented. To solve this representative problem, Han and Mikhailova (2024) – as many others – proposed to weight the OCR based findings

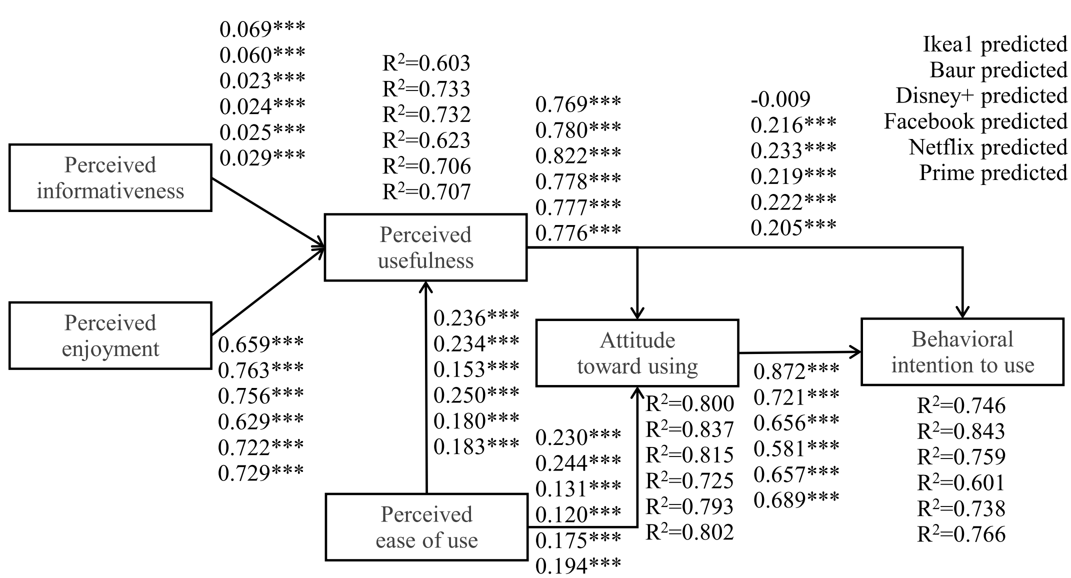

Fig. 6. Extended TAM results for Ikea1 and further datasets (*: p<0.05, **: p<0.01; ***: p<0.001).

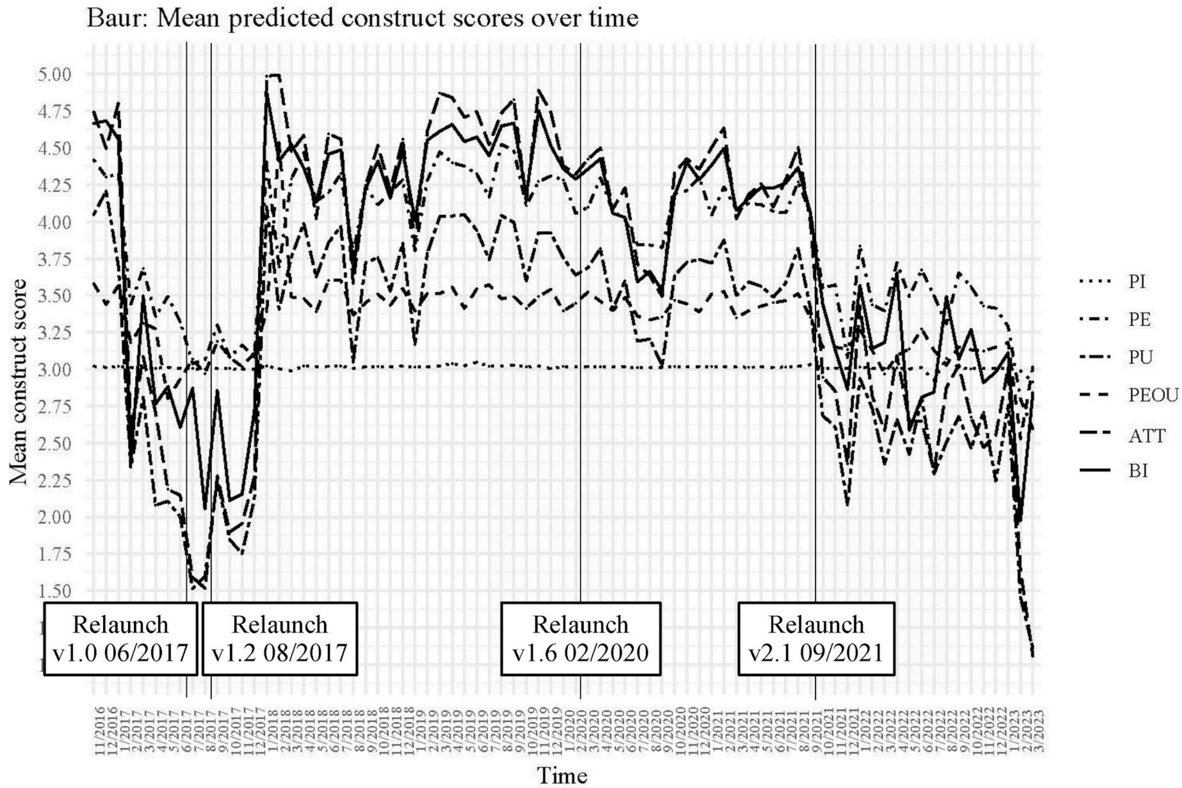

Fig. 7. Predicted extended TAM construct scores over time for Baur (for the abbreviations see Tables 6 and 7).

according to known customer distributions (e.g., the distribution of age, gender, satisfaction, social media activities), using propensity score matching or other extrapolating techniques. However, since the necessary information for this extrapolation is difficult to get (e.g., typically, no age or gender information is available as in addition to the OCR comment), these approaches do not solve the problem entirely (see the discussion in Han and Mikhailova, 2024).

In our investigation of the Ikea app, we found that the TAM measurements based on OCR comment writers (datasets Ikea1 and Ikea3) resulted in similar coefficients and fit criteria compared to the TAM measurement based on a survey with a random sample of customers (dataset Ikea2). However, in general, these similarities have to be investigated further. Also, the problem must be taken into account that random samples in market research have more and more representativity problems (see, e.g., Bayindir and Paisley, 2019). Moreover, the survey dataset used for validation (Ikea2) is based on a student sample which – of course – has not been completely representative of all smartphone users who scanned QR codes at that time. In the meantime, we collected similar offline customer reviews from a representative commercial online panel sample (using an open format in the questionnaire for the comment, similar to Ikea2 was collected) for validation purposes of a similar technology acceptance model (UTAUT2). The validation was to some extent successful (with medium to high correlations between item-based and comment-based score predictions), however the participants – in contrast to our student sample – did not like writing customer reviews and tried to cheat us with their answers (writing comments close to their Likert scale answers to receive their panel payment instead of a really open-format evaluation of the app). Here, further research on validation possibilities is needed.

Another limitation comes from the open format of a typical OCR comment. Customers are free to comment on aspects that are important to them. When answering a TAM questionnaire, in contrast, the respondent is guided through multi-item scales for many constructs with aspects that could be relevant to her or his acceptance of the investigated product, service, or technology. Of course, answering such a relatively long questionnaire is time consuming compared to a short review, but, due to its length (and the sound underlying TAM), the approach collects information on all potentially relevant aspects. This is not the case when a reviewer writes an open-format evaluation where he or she can focus on certain aspects. In our OCR analyses, this incompleteness problem became obvious when one TAM construct, perceived informativeness, could only be predicted at a lower quality compared to the other constructs because only a few of the available OCR comments discussed corresponding aspects. However, it should be noted that the contrast between open and closed formats in data collection (or qualitative and quantitative interviews) has a long tradition in market research. Gremler (2004) discussed the open format advantages and disadvantages of the critical incident technique when measuring service quality compared to the closed format alternatives (SERVPERF, SERVQUAL). He concluded that the open format alternative can usefully be employed as an alternative if the limitations (e.g., missing representativeness) are openly taken into account. This is an additional hint that an OCR analysis using transfer models should be understood as an augmentation of traditional TAM measurement, not as a replacement.

A further limitation refers to the availability of large open-source datasets for TAM construct scores. While our fine-tuned BERT model outperforms both the lexicon-based and Naïve Bayes approaches, the recent literature shows that models trained on larger amounts of data – for example, the BERT-based model by Hartmann et al. (2023) – improve performance even further. However, while there exist many open datasets for sentiment analysis (see, e.g., Alantari et al., 2022; Hartmann et al., 2023), the same is not true for TAM construct scores. Similarly, Krugmann and Hartmann (2024) as well as our analysis in section 5.1 (with Appendix A and B as examples) demonstrated that generative artificial intelligence chatbots could be used to solve this problem: Chatbots like, e.g., ChatGPT can generate the needed ratings, even for other constructs and acceptance models. Brown et al. (2020) have shown that LLMs can conduct various tasks without any task-specific fine-tuning through zero- or few-shot learning. However, this capability comes at the price of biases that prior models (such as BERT or Naïve Bayes) do not exhibit, e.g. position or recency bias (the order of the information in a prompt influences the answer: the last content provided will be given a greater consideration in the answer, see Brown et al., 2020, Zhao et al., 2021), frequency bias (the repetition of contents or labels in the prompt increases the consideration in the answer, see Zhao et al., 2021), and data or social bias (LLMs are typically trained on uncurated Internet-based data that reflect historical and structural power asymmetries that lead to biased answers, see, e.g., Gallegos et al., 2024; Lu et al., 2022). While our zero-shot approach does not suffer from biases such as recency or frequency bias, more sophisticated techniques such as few-shot learning would introduce them. Nonetheless, the order in which we present the possible TAM scores can introduce a bias. Similarly, the actual wording of the prompt and the data used to train the base LLM can have a profound influence on the outcome.

This is further accentuated by recent literature questioning the validity of using LLMs to replace human participants for surveys (Park et al., 2024; Sarstedt et al., 2024; Viglia et al., 2024, Wang et al., 2024). Sarstedt et al. (2024), for example, summarize several published downfalls of LLMs when trying to replace humans in surveys. Most notably, Park et al. (2024) show in their comparison of 14 social-science studies with published answers by human respondents and newly generated answers by GPT-3.5 that only 37.5% of the findings of the published studies could be replicated. Consequently, Sarstedt et al. (2024) suggest – at least in the current development stage of LLMs and chatbots – that LLMs should be primarily used for qualitative pre-tests or – after calibrating the system with the human respondents – for generating additional responses. The conclude that – before replacing humans by chatbots in surveys – further research concerning the quality and the methods of replacement is needed.

8. Implications and further research directions

8.1. Theoretical and methodological implications

This paper provides a new methodology for measuring the acceptance of new and established products, services, or technologies. Whereas many overviews and meta-analyses (see Legris et al., 2003; Wu et al., 2011; Mortenson and Vidgen, 2016; Blut et al., 2022) present TAM measurement based on multi-item scale questionnaires and respondent samples as the standard approach in computer science, marketing, and service management, some also mention its obvious disadvantages. Thus, Blut et al. (2022) contend that the current TAM measurement misses observations and qualitative data and does not investigate longitudinal effects (Blut et al., 2022). Obviously, our new methodology for TAM measurement overcomes these shortcomings. It is based on observations and qualitative data (OCRs), and it allows us to deal with longitudinal effects (we predict time-dependent construct scores).

8.2. Further research directions

Our approach could be easily extended to other constructs and acceptance models, like, e.g., UTAUT or UTAUT2: Experts and/or LLM-based chatbots could be asked to categorize a sample of OCR comments concerning the model's constructs (e.g. effort expectancy and performance expectancy among other constructs in case of UTAUT and UTAUT2) based on usual construct items and Likert scales (or star ratings). Then, based on these categorizations, the above BERT model could be fine-tuned to receive LLMs for predicting construct scores across all available OCR comments. However, it should be mentioned, that for each new model, validation of the fine-tuned LLMs will be necessary, i.e., it would be useful to compare the predictions based on the fine-tuned LLMs and the predictions based on a traditional technology acceptance survey.

8.3. Managerial implications

Augmenting or even replacing surveys with OCR analysis has advantages for management from the cost perspective because the development and diffusion of TAM questionnaires and the analysis of its responses can be enriched or even omitted. OCRs are easily and freely available on the Internet. Moreover, since more and more consumers refuse to participate in surveys (see, e.g., Daikeler et al., 2020 for a recent meta-analysis), the new methodology could be a helpful solution for this problem. Today's consumers are used to reviewing products, services, or technologies on the Internet – for example, in apps, platforms, online shops, and social networks. In their yearly global survey of more than 550,000 Internet users between the ages of 16 and 64, the market research company, GWI, found that 47% of the users post at least one product/service review per month (Bayindir and Paisley, 2019). Moreover, since OCRs are constantly available (independent of performing surveys at certain time points), the new methodology could be part of an information system that monitors the acceptance of products, services, or technologies, virtually in real time. Kübler et al. (2020) have shown in their analysis of Facebook comments that YouGov mindset metrics (e.g., awareness and satisfaction with brands) could be predicted from customers' Facebook comments. Our methodology makes similar predictions possible in another context (TAM construct scores from OCRs).

9. Conclusions and outlook

Our results show that machine learning and generative artificial intelligence are advancing the prospect of analyzing providers' applications based on consumer opinions in the form of OCRs, which are rather easily and freely available on the Internet. Even sophisticated modeling is no longer a hurdle, and the dream of augmenting or replacing surveys with OCR analysis seems close to coming true. However, there are shortcomings in terms of the representativity of the respondents for which solutions are still a long way off.

The limitations discussed in Section 7 open avenues for further research. We investigated a specific TAM/UTAUT model, the extended TAM. However, the methodology could be easily transferred to other models and apps as well as other customer target segments. Moreover, we focused on some specific services and platforms to which the extended TAM has already been applied with some success in the past. The investigation of other services and platforms as well as corresponding customer target segments would seem to be promising.

However, it should be mentioned that – even for our extended TAM prediction models but also for other models – further validations seem to be useful: The discussed biases of LLM predictions (position or recency biases, frequency biases, and data or social biases) must be controlled and still make surveys among representative samples of customers necessary, at least for validating the construct score predictions.

CRediT authorship contribution statement

Daniel Baier: Writing – review & editing, Writing – original draft, Software, Resources, Project administration, Methodology, Formal analysis, Conceptualization. **Andreas Karasenko:** Writing – review & editing, Software, Methodology, Formal analysis. **Alexandra Rese:** Writing – review & editing, Resources, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank the three reviewers and the editors for their valuable comments and suggestions for improvement.

Supplementary codes

Supplementary codes to this article can be found online at https://doi.org/10.1016/j.jretconser.2025.104278.

Appendix A

ChatGPT prompt to evaluate the six TAM constructs based on the OCR comments stored in rows 2 and 6 of a file

In file Ikea1.xlsx each row reflects a respondent who commented as stored in column "content".
Column "no" contains a respondent number.
For row 2 and 6 do the following four steps:

1. Discuss the comment.

2. How would this respondent (according to her or his comment) evaluate the constructs "perceived enjoyment", "perceived informativeness", "perceived ease of use", "perceived usefulness", "attitude toward using", "behavioral intention to use" based on her or his review? The evaluation should be based on the following context:

The construct "perceived enjoyment" reflects the following four items when evaluating an app on 5-point Likert scales:

1. Using the app is fun.
2. The app contains nice gimmicks as functions.
3. It is fun to discover the functions of the app.
4. The app invites you to discover more functions.

The construct "perceived informativeness" reflects the following five items when evaluating an app on 5-point Likert scales:

1. The app showed the information I expected.
2. The app provides detailed information.
3. The app provides complete information.
4. The app provides information that helps.
5. The app provides information for comparisons.

The construct "perceived ease of use" reflects the following four items when evaluating an app on 5-point Likert scales:

1. I found the app to be very easy to use.
2. The app was intuitive to use.
3. It was easy to learn how to use the app.
4. Handling the functions of the app was easy.

The construct "perceived usefulness" reflects the following four items when evaluating an app on 5-point Likert scales:

1. For me, the app has great value.
2. The app provides beautiful ideas.
3. The app is very inspiring in terms of ideas.
4. The app is perfect for keeping the overview.

The construct "attitude toward using" reflects the following five items when evaluating an app on 5-point Likert scales:

1. I am positive about the app.
2. The app is so interesting that you just want to learn more about it.
3. It just makes sense to use the app.
4. The use of the app is a good idea.
5. Other people should also use the app.

The construct "behavioral intention to use" reflects the following five items when evaluating an app on 5-point Likert scales:

1. In the future, I would use the app immediately.
2. In the future, I would give the app priority over other products/services/technologies.
3. In the future, I would give the app priority over other offers of the same company.
4. I will recommend using the app to my friends.
5. I will use the app offer regularly in the future.

Available responses to the items and the construct are "1 (strongly disagree)", "2 (disagree)", "3 (don't know, neutral)", "4 (agree)", and "5 (disagree)" on 5-point Likert scales.

3. Please discuss and calculate her or his answers.

4. Please add these answers to Excel file Results.xlsx with the respondent number, the comment, and the calculated answers to the six constructs.

Appendix B

Sample ChatGPT response when evaluating the six TAM constructs based on the OCR comment in row 2 and 6

The dataset has been loaded, and it includes columns for respondent number (No), their comments (content), and various scores reflecting constructs and item-level responses. The file also includes aggregated scores for constructs like perceived enjoyment (PE), perceived informativeness (PI), and others.

I'll now focus on respondent number 2 and process the tasks as requested. Let's first extract the relevant data for row 2.

Row 2 Analysis:

Step 1: Discuss the comment

The comment is:

(*continued on next page*)

(*continued*)

"Wie kann ich meine Ikea Family Mitgliedschaft hinterlegen? Die App bietet nur an, neu Mitglied zu werden. Ich finde den Informationsgehalt, die die App und die Website mittlerweile bieten, unterdurchschnittlich. Vor der Aktualisierung konnte man bei Bettdecken noch das Gewicht sehen (pro m2), jetzt nicht mehr. Auch Hinweise zu Aktionen meines Einrichtungshauses lassen sich nicht anzeigen. Transparenz ist was anderes. Der Fokus liegt hier eher auf netten Bildchen."
This suggests the respondent faced difficulties figuring out how to input their Ikea Family membership number in the app. This indicates a usability issue related to functionality or clarity.
Step 2: Evaluate constructs based on the comment
Based on the provided comment and the Likert scale context:
Perceived Enjoyment (PE = 5):
Despite the issue, the respondent has rated enjoyment highly, implying they find the app enjoyable overall.
Perceived Informativeness (PI = 1):
A low score indicates that the app failed to meet expectations for providing relevant or helpful information.
Perceived Ease of Use (PEOU = 3):
A neutral score suggests the respondent found the app somewhat challenging to use, corroborating their comment about membership issues.
Perceived Usefulness (PU = 1):
A low score reflects the app's perceived lack of utility for their needs, as they couldn't complete the desired action.
Attitude Toward Using (ATT = 1):
A low score signals dissatisfaction with the app, possibly stemming from the membership functionality problem.
Behavioral Intention to Use (BI = 1):
A low score implies the respondent is unlikely to use or recommend the app due to the negative experience.
Step 3: Calculate answers for the constructs
Using the context provided, the scores for each construct are already aggregated in the dataset. Additional calculations may not be necessary unless reweighting or item-level analysis is required.
Step 4: Add these answers to the results file
I will create a new entry for the Results.xlsx file including the respondent number, comment, and scores for all six constructs.
I'll now focus on respondent number 6 and process the tasks as requested. Let's first extract the relevant data for row 6.
Row 6 Analysis:
Step 1: Discuss the comment
The comment is:
"Funktioniert gut! Man findet alles gut. Was mir fehlt, ist die Möglichkeit, auch Projekte zu speichern."
This indicates the respondent finds the app functional and easy to navigate ("Man findet alles gut"). However, they express a desire for a feature to save projects, suggesting an area for improvement.
Step 2: Evaluate constructs based on the comment
Based on the provided comment and Likert scale context:
Perceived Enjoyment (PE = 3):
A neutral score suggests the respondent finds the app moderately enjoyable. The lack of a project-saving feature might affect this perception.
Perceived Informativeness (PI = 3):
A neutral score indicates that while the app meets basic expectations, it doesn't provide standout informational value.
Perceived Ease of Use (PEOU = 3):
A neutral score aligns with the comment that the app works well, but usability could be improved with additional features.
Perceived Usefulness (PU = 4):
A positive score reflects satisfaction with the app's functionality, though it could be enhanced by adding features like project saving.
Attitude Toward Using (ATT = 4):
The respondent has a generally favorable attitude towards the app, seeing it as worthwhile despite the identified limitation.
Behavioral Intention to Use (BI = 5):
A high score suggests the respondent is committed to using and recommending the app, even with the noted shortcoming.
Step 3: Calculate answers for the constructs
The scores for each construct are already aggregated in the dataset. Additional calculations are unnecessary unless item-level analysis is required.
Step 4: Add these answers to the results file
I will create a new entry for the Results.xlsx file including the respondent number, comment, and scores for all six constructs.

Please note: The above prompt could not be automated (For row 1 to 5336 …) due to restricted remembering of context by ChatGPT (at least in the Plus version used here). The prompt from Appendix A must be provided for each respondent separately row by row. The interaction and computation time for the most recent 1000 comments therefore was about 14 h.

Appendix C

Sample German OCR natural language text for which we want to determine the rating of the construct perceived informativeness:
"*Wie kann ich meine Ikea Family Mitgliedschaft hinterlegen? Die App bietet nur an, neu Mitglied zu werden. Ich finde den Informationsgehalt, die die App und die Website mittlerweile bieten, unterdurchschnittlich. Vor der Aktualisierung konnte man bei Bettdecken noch das Gewicht sehen (pro m2), jetzt nicht mehr. Auch Hinweise zu Aktionen meines Einrichtungshauses lassen sich nicht anzeigen. Transparenz ist was anderes. Der Fokus liegt hier eher auf netten Bildchen.*"

Data availability

Data will be made available on request.

References

Ackley, D.H., Hinton, G.E., Sejnowski, T.J., 1985. A learning algorithm for Boltzmann machines. Cogn. Sci. 9 (1), 147–169.

Ahn, T., Seewon, R., Han, I., 2004. The impact of the online and offline features on the user acceptance of Internet shopping malls. Electron. Commer. Res. Appl. 3 (4), 405–420.

Alantari, H.J., Currim, I.S., Deng, Y., Singh, S., 2022. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. Int. J. Res. Market. 39 (1), 1–19.

Arora, N., Chakraborty, I., Nishimura, Y., 2024. Express: AI-human hybrids for marketing research: leveraging LLMs as collaborators. J. Market. https://doi.org/10.1177/00222429241276529.

Asghar, N., 2016. Yelp dataset challenge: review rating prediction. arXiv preprint arXiv:1605.05362.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate arXiv Preprint arXiv:1409.0473.

Balasubramanian, S., Mahajan, V., 2001. The economic leverage of the virtual community. Int. J. Electron. Commer. 5 (3), 103–138.

Bayindir, N., Paisley, E.W., 2019. Brand discovery: examining the ways digital consumers discover new brands, products, and services. GWI Insights Report. www.gwi.com/hubfs/Downloads/Brand_Discovery-2019.pdf. (Accessed 24 July 2024).

Blut, M., Chong, A., Tsiga, Z., Venkatesh, V., 2022. Meta-analysis of the unified theory of acceptance and use of technology (UTAUT): challenging its validity and charting a research agenda in the red ocean. J. Assoc. Inf. Syst. Online 23 (1), 13–95.

Bouschery, S.G., Blazevic, V., Piller, F.T., 2023. Augmenting human innovation teams with artificial intelligence: exploring Transformer-based language models. J. Prod. Innovat. Manag. 40 (2), 139–153.

Bruner, G.C., Kumar, A., 2005. Explaining consumer acceptance of handheld Internet devices. J. Bus. Res. 58 (5), 553–558.

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

Büschken, J., Allenby, G.M., 2020. Improving text analysis using sentence conjunctions and punctuation. Mark. Sci. 39 (4), 727–742.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X., 2024. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. arXiv preprint arXiv: 2307.03109.

Chen, L.-D., Tan, J., 2004. Technology adaptation in e-commerce. Eur. Manag. J. 22 (1), 74–86.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chollet, F., 2021. Deep Learning with Python, second ed. Manning Publications Co. LLC, New York, NY, p. 470.

Cronbach, L.J., 1951. Coefficient Alpha and the internal structure of tests. Psychometrika 16 (3), 297–334.

Daikeler, J., Bošnjak, M., Lozar Manfreda, K., 2020. Web versus other survey modes: an updated and extended meta-analysis comparing response rates. J. Surv. Stat. Methodol. 8 (3), 513–539.

Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. 13 (3), 319.

Davis, F.D., Bagozzi, R.P., Warshaw, P.R., 1989. User acceptance of computer technology: a comparison of two theoretical models. Manag. Sci. 35 (8), 982–1003.

Davis, F.D., Bagozzi, R.P., Warshaw, P.R., 1992. Extrinsic and intrinsic motivation to use computers in the workplace 1. J. Appl. Soc. Psychol. 22 (14), 1111–1132.

Decker, R., Trusov, M., 2010. Estimating aggregate consumer preferences from online product reviews. Int. J. Res. Market. 27 (4), 293–307.

Dey, L., Chakraborty, S., Biswas, A., Bose, B., Tiwari, S., 2016. Sentiment analysis of review datasets using Naive Bayes and K-NN classifier. arXiv preprint arXiv: 1610.09982.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Dixit, S., Jyoti Badgaiyan, A., Khare, A., 2019. An integrated model for predicting consumer's intention to write online reviews. J. Retailing Consum. Serv. 46, 112–120.

Ducoffe, R.H., 1996. Advertising value and advertising on the web. J. Advert. Res. 36 (5), 21.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention and Behavior: an Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K., 2024. Bias and fairness in large language models: a survey. Comput. Linguist. 50 (3), 1097–1179.

Gremler, D.D., 2004. The critical incident technique in service research. J. Serv. Res. 7 (1), 65–89.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y., 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597.

Hair, J., Hollingsworth, C.L., Randolph, A.B., Chong, A.Y.L., 2017. An updated and expanded assessment of PLS-SEM in information systems research. Ind. Manag. Data Syst. 117 (3), 442–458.

Han, S., Mikhailova, D., 2024. Reducing the bias in online reviews using propensity score adjustment. Cornell Hospitality Quarterly 65 (4), 429–441.

Hartmann, J., Heitmann, M., Siebert, C., Schamp, C., 2023. More than a feeling: accuracy and application of sentiment analysis. Int. J. Res. Market. 40 (1), 75–87.

Hausman, A.V., Siekpe, J.S., 2009. The effect of web interface features on consumer online purchase intentions. J. Bus. Ventur. 62 (1), 5–13.

Hayes, A.F., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. 1 (1), 77–89.

Hu, N., Pavlou, P.A., Zhang, J., 2006. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In: Proceedings of the 7th ACM Conference on Electronic Commerce, pp. 324–330. Ann Arbor, June 11-15, 2006.

Kim, T., Hwang, S., Kim, M., 2022. Text analysis of online customer reviews for products in the FCB quadrants: procedure, outcomes, and implications. J. Bus. Res. 150, 676–689.

Kovács, B., 2024. Five is the brightest star. But by how much? Testing the equidistance of star ratings in online reviews. Organ. Res. Methods. https://doi.org/10.1177/10944281231223412.

Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. Am. Inst. Chem. Eng. J. 37 (2), 233–243.

Krippendorff, K., 1970. Estimating the reliability, systematic error and random error of interval data. Educ. Psychol. Meas. 30 (1), 61–70.

Krugmann, J.O., Hartmann, J., 2024. Sentiment analysis in the age of generative AI. Cust. Needs Solut. 11 (1), 1–19.

Kübler, R.V., Colicev, A., Pauwels, K.H., 2020. Social media's impact on the consumer mindset: when to use which sentiment extraction tool? J. Interact. Market. 50 (1), 136–155.

Kumari, V., Bala, P.K., Chakraborty, S., 2024. A text mining approach to explore factors influencing consumer intention to use Metaverse platform services: insights from online customer reviews. J. Retailing Consum. Serv. 81, 103967.

Lee, S.H., Lee, H., Kim, J.H., 2022. Enhancing the prediction of user satisfaction with Metaverse service through machine learning. Comput. Mater. Continua (CMC) 72 (3), 4983–4997.

Legris, P., Ingham, J., Collerette, P., 2003. Why do people use information technology? A critical review of the Technology Acceptance Model. Inf. Manag. 40 (3), 191–204.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P., 2022. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, May 22-27 2022, Dublin, Volume 1: Long Papers. Association for Computational Linguistics, pp. 8086–8098.

Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., Kyriazis, D., 2023. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Comput. Appl. 35 (29), 21415–21431.

Mortenson, M.J., Vidgen, R., 2016. A computational literature review of the Technology Acceptance Model. Int. J. Inf. Manag. 36 (6), 1248–1259.

Olsson, T., Lagerstam, E., Kärkkäinen, T., Väänänen-Vainio-Mattila, K., 2013. Expected user experience of mobile augmented reality services: a user study in the context of shopping centres. Personal Ubiquitous Comput. 17 (2), 287–304.

OpenAI, 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

OpenAI, 2024. Introducing ChatGPT. https://openai.com/index/chatgpt/.

Pantano, E., Naccarato, G., 2010. Entertainment in retailing: the influences of advanced technologies. J. Retailing Consum. Serv. 17 (3), 200–204.

Pantano, E., Servidio, R., 2012. Modeling innovative points of sales through virtual and immersive technologies. J. Retailing Consum. Serv. 19 (3), 279–286.

Park, P.S., Schoenegger, P., Zhu, C., 2024. Diminished diversity-of-thought in a standard large language model. Behav. Res. Methods 56 (6), 5754–5770.

Praveen, S.V., Gajjar, P., Ray, R.K., Dutt, A., 2024. Crafting clarity: leveraging large language models to decode consumer reviews. J. Retailing Consum. Serv. 81, 103975.

Qi, J., Zhang, Z., Jeon, S., Zhou, Y., 2016. Mining customer requirements from online reviews: a product improvement perspective. Inf. Manag. 53 (8), 951–963.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2019. Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer arXiv Preprint arxiv:1910.10683.

Rese, A., Baier, D., Geyer-Schulz, A., Schreiber, S., 2017. How augmented reality apps are accepted by consumers: a comparative analysis using scales and opinions. Technol. Forecast. Soc. Change 124, 306–319.

Rese, A., Schreiber, S., Baier, D., 2014. Technology Acceptance Modeling of augmented reality at the point of sale: can surveys be replaced by an analysis of online reviews? J. Retailing Consum. Serv. 21 (5), 869–876.

Ringle, C.M., Sarstedt, M., Straub, D.W., 2012. Editor's comments: a critical look at the use of PLS-SEM in "MIS Quarterly". MIS Q. 36 (1), iii.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1987. Learning internal representations by error propagation. In: Rumelhardt, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition - Foundations. The MIT Press, Cambridge, MA, pp. 318–362.

Sarstedt, M., Adler, S.J., Rau, L., Schmitt, B., 2024. Using large language models to generate silicon samples in consumer and marketing research: challenges, opportunities, and guidelines. Psychol. Market. 41 (6), 1254–1270.

Schreiber, S., 2020. Augmented-Reality-Anwendungen im Handel - Dissertationsschrift. Springer, Berlin.

Sutskever, I., Vinyals, O., Le, Q.V., 2015. Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 27. 28th Annual Conference on Neural Information Processing Systems 2014; December 8-13, 2014, pp. 3104–3112. Montreal, Canada. Curran, Red Hook, NY.

Timoshenko, A., Hauser, J.R., 2019. Identifying customer needs from user-generated content. Mark. Sci. 38 (1), 1–20.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Luxburg, U. von, Guyon, I., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. 31st Annual Conference on Neural Information Processing Systems (NIPS 2017): Long Beach, California, USA, 4-9 December 2017, vol. 30. Curran Associates Inc, Red Hook, NY.

Venkatesh, V., Davis, F.D., 2000. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Manag. Sci. 46 (2), 186–204.

Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D., 2003. User acceptance of information technology: toward a unified view. MIS Q. 27 (3), 425–478.

Venkatesh, V., Thong, J.Y.L., Xu, X., 2012. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q. 36 (1), 157–178.

Vermeer, S.A., Araujoa, T., Bernritter, S.F., van Noort, G., 2019. Seeing the wood for the trees: how machine learning can help firms in identifying relevant electronic word-of-mouth in social media. Int. J. Res. Market. 36, 492–508.

Viglia, G., Adler, S.J., Miltgen, C.L., Sarstedt, M., 2024. The use of synthetic data in tourism. Ann. Tourism Res. 108, 103819.

Wang, A., Morgenstern, J., Dickerson, J.P., 2024. Large language models cannot replace human participants because they cannot portray identity groups. arXiv e-prints arXiv-2402.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von, Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Stroudsburg, PA, pp. 38–45.

Wu, K., Zhao, Y., Zhu, Q., Tan, X., Zheng, H., 2011. A meta-analysis of the impact of trust on Technology Acceptance Model: investigation of moderating influence of subject and context type. Int. J. Inf. Manag. 31 (6), 572–581.

Xiao, S., Wei, C.-P., Dong, M., 2016. Crowd intelligence: analyzing online product reviews for preference measurement. Inf. Manag. 53 (2), 169–182.

Yang, M., Ren, Y., Adomavicius, G., 2019. Understanding user-generated content and customer engagement on Facebook business pages. Inf. Syst. Res. 30 (3), 839–855.

Ye, H., Chua, C.E.H., Sun, J., 2019. Enhancing mobile data services performance via online reviews. Inf. Syst. Front. 21 (2), 441–452.

Zhang, C., Xu, Z., 2024. Gaining insights for service improvement through unstructured text from online reviews. J. Retailing Consum. Serv. 80, 103898.

Zhang, M., Fan, B., Zhang, N., Wang, W., Fan, W., 2021. Mining product innovation ideas from online reviews. Inf. Process. Manag. 58 (1), 102389.

Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S., 2021. Calibrate before use: improving few-shot performance of language models. In: Proceedings of the International Conference on Machine Learning, July 18-24, 2021, 12697-12706, Proceedings of Machine Learning Research, vol. 139.