Baier, Daniel; Decker, Reinhold; Asenova, Yana

Article

Collecting and analyzing user-generated content for decision support in marketing management: An overview of methods and use cases

Schmalenbach Journal of Business Research (SBUR)

Provided in Cooperation with:
Schmalenbach-Gesellschaft für Betriebswirtschaft e.V.

Suggested Citation: Baier, Daniel; Decker, Reinhold; Asenova, Yana (2025) : Collecting and analyzing user-generated content for decision support in marketing management: An overview of methods and use cases, Schmalenbach Journal of Business Research (SBUR), ISSN 2366-6153, Springer, Heidelberg, Vol. 77, Iss. 3, pp. 419-455, https://doi.org/10.1007/s41471-025-00208-7

This Version is available at:
https://hdl.handle.net/10419/331932

Schmalenbach Journal of Business Research (2025) 77:419–455
https://doi.org/10.1007/s41471-025-00208-7

ORIGINAL ARTICLE

Collecting and Analyzing User-Generated Content for Decision Support in Marketing Management: An Overview of Methods and Use Cases

Daniel Baier · Reinhold Decker · Yana Asenova

Received: 3 October 2024 / Accepted: 11 February 2025 / Published online: 12 March 2025

Abstract User-generated content (UGC) is generally understood as an expression of opinion in many forms (e.g., complaints, online customer reviews, posts, testimonials) and data types (e.g., text, image, audio, video, or a combination thereof) that has been created and made available by users of websites, platforms, and apps on the Internet. In the digital age, huge amounts of UGC are available. Since UGC often reflects evaluations of brands, products, services, and technologies, many consumers rely on UGC to support and secure their purchasing and/or usage decisions. But UGC also has significant value for marketing managers. UGC allows them to easily gain insights into consumer attitudes, preferences, and behaviors. In this article, we review the literature on UGC-based decision support from this managerial perspective and look closely at relevant methods. In particular, we discuss how to collect and analyze various types of UGC from websites, platforms, and apps. Traditional data analysis and machine learning based on feature extraction methods as well as discriminative and generative deep learning methods are discussed. Selected use cases across various marketing management decision areas (such as customer/market selection, brand management, product/service quality management, new product/service development) are summarized. We provide researchers and practitioners with a comprehensive understanding of the current state of UGC data collection and analysis and help them to leverage this powerful resource effectively.

✉ Daniel Baier
Chair of Marketing & Innovation, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany
E-Mail: daniel.baier@uni-bayreuth.de

Reinhold Decker
Chair of Marketing, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany
E-Mail: rdecker@uni-bielefeld.de

Yana Asenova
Kühne Logistics University, Großer Grasbrook 17, 20457 Hamburg, Germany
E-Mail: yana.asenova@klu.org

Moreover, we shed light on potential applications in managerial decision support and identify research questions for further exploration.

Keywords User-generated content (UGC) · Online customer reviews (OCRs) · Posted images and videos · Web scraping · Machine learning · Discriminative deep learning · Generative deep learning

1 Introduction

Data-based decision support has a long tradition in marketing (for a comprehensive review see Wedel and Kannan 2016). As early as the 1920s and 1930s, consumer product providers, such as Procter & Gamble, and market research companies, such as A. C. Nielsen, Burke, and GfK, collected consumer attitudes, preferences, and sales data to support marketing decisions. Since then, the collection and analysis of external and internal consumer data has become widespread, allowing product/service providers to gain consumer insights and to support decisions, often supported by advanced methods from academic research (Wedel and Kannan 2016). Roberts et al. (2014) investigated the impact of this data-driven approach on marketing practice and concluded from surveys among managers and citation analyses that the following decision areas have benefited most from this approach (with decreasing importance): pricing management, customer/market selection, marketing strategy (product line, multi-product, portfolio decisions), new product/service management, relationship management (customer value assessment and maximization, customer acquisition and retention decisions), sales force management, product/service quality management, and brand management.

Today, in the digital age, it is even easier to support marketing decisions based on consumer data since consumers are used to publish their attitudes, preferences, and purchase/usage intentions on the Internet (see, e.g., Dorner et al. 2020). Blogs, company websites, online shops, platforms, review sites, and social networks offer a multitude of possibilities to comment on brands, products/services, and technologies. Consumers make extensive use of these possibilities. So, for instance, in their yearly global survey of more than 550,000 Internet users aged 16–64, the market research company GWI discovered that 47% of the users post at least one online customer review (OCR) per month (Bayindir and Paisley 2019). The online retailer Amazon recently reported that approximately 125 million of its customers contribute nearly 1.5 billion OCRs per year (Schermerhorn 2023). It was reported that more than a third of all Internet users worldwide consider OCRs reliable and assign them an influence on their purchasing and/or usage decisions (Bayindir and Paisley 2019). According to a recent survey by the market research company BrightLocal, Google's OCRs of companies help 83% of Internet users when selecting adequate nearby offers (Paget 2024).

For companies, these huge amounts of so-called user-generated content (UGC) are an attractive data source for insight generation compared to survey data and sales data. UGC often better reflects attitudes, preferences, and satisfaction than sales data. In contrast to surveys, UGC is freely and easily available, virtually in

real-time, and does not require questionnaire development and distribution (Timoshenko and Hauser 2019). The review website Trustpilot reflects the attractiveness UGC has for companies by reporting that more than one million companies have registered for its premium service with extensive opportunities to access OCRs, to stimulate customers to post OCRs, and to generate advanced insights (Brooke 2024). This attractiveness is also reflected in many academic research articles that discuss methods to analyze UGC and how it is utilized by companies. So, for example, Li et al. (2022) provide a literature review in which they found that currently more than 3390 academic research articles have been published on methods for UGC analysis in e-commerce, with numbers steadily increasing since 2013. The thematic analysis, among others, revealed that sentiment analysis and user preference mining have received the most attention from researchers, with consumer profiling and product design being the dominant applications.

In this article, we review the widespread utilization of UGC for marketing decision support in more detail. Section 2 defines UGC, Sect. 3 focuses on UGC data collection, and Sect. 4 deals with methods for UGC analysis. Section 5 gives an overview of published use cases. Section 6 develops a UGC utilization guideline for researchers and managers, based on these literature reviews. Finally, Sect. 7 presents conclusions and an outlook on future research topics.

2 UGC

UGC, alternatively known as user-created content, is a term that often refers to the well-known OECD (Organization of Economic Co-operation and Development) definition by Vickery and Wunsch-Vincent (2007) in a report on the participative web (Naab and Sehl 2017). Their definition reflects the trend that the Internet is more and more characterized by the participation of and the interaction between users. In addition to their own, self-created content (e.g., product/service descriptions, navigation helps, and news created by employees), websites, platforms, and apps provide users with the means to produce, customize, and (co-)develop content in many forms (e.g., OCRs, improvement proposals, testimonials) and with many data types (e.g., text, image, audio, video, or a combination thereof). According to this OECD definition, content has to meet the following three criteria simultaneously to be classified as UGC:

- The content has been published: It is made available ("posted") by the user on a publicly accessible website, platform, or app or at least a website, platform, or app that can be accessed by a larger group of users (e.g., registered members of a social network). This criterion excludes bilateral information exchange between users or between a user and a company (e.g., via mail or instant message).
- The content reflects a degree of creative effort by the user. The posted information consists of text, image, audio, and/or video in a context that expresses her or his opinions on oneself, other people, objects, or organizations. Text, image, audio, and video that are simply uploaded for storing purposes or as simple expressions

without particular reason or specific context are excluded by this criterion (Naab and Sehl 2017).
- The content is created outside professional routines and practices. This criterion excludes content produced by the employees of a company (e.g., product/service details in an online shop).

Typical websites, platforms, and apps that support content generation by users are the following (Naab and Sehl 2017):

- (Internet) Forums and blogs are websites created by individuals or organizations where users can post comments, usually about pre-defined topics. Forum or blog hosting platforms, such as WordPress (www.uniteddomains.com) or Medium (www.medium.com), support the development and hosting of these websites.
- Product/service provider websites and apps (e.g., online shops, such as www.adidas.com, www.ikea.com, www.lidl.com, www.mediamarkt.de), or platforms (such as www.amazon.com and www.otto.de) allow users to purchase products/services from a single company or multiple companies but also to post OCRs.
- Review platforms, such as www.trustpilot.com, www.provenexpert.de, www.reviews.io, and www.yelp.com, are independent of the product/service supply chain and also allow users to post OCRs.
- Wikis, such as Wikipedia (www.wikipedia.com) and Fandom (www.fandom.com), are encyclopedias whose content is co-developed by their users.
- Social networks, such as Facebook, Instagram, Snapchat, and X, are platforms that allow users to upload content, to construct a narrative of their life, or to interact with others (e.g., by commenting or liking).
- Media hosting platforms and apps, such as Flickr, TikTok, Vimeo, and YouTube, allow images and videos to be posted but also permit users to comment on posts.

Due to its characteristics and across these websites, platforms, and apps, UGC is a valuable source of information for marketing managers. For example, Yang et al. (2019) found that UGC often contains customer complaints and suggestions that could be a starting point for product/service quality improvement. The authors investigated customer posts on Facebook business pages of Fortune 500 companies and applied grounded theory to derive typical UGC content characteristics. Two research assistants coded a small sample of customer posts and identified altogether seven UGC content categories: “positive testimonial and appreciation”, “complaint about product/service quality”, “complaint about money issues”, “complaint about social and environmental issues”, “customer questions”, “customer suggestions”, and “others”. Then, a large number of MTurk workers characterized all available customer posts. The respondents received a list of the seven UGC content categories and were asked to characterize the presented customer posts accordingly. Across the seven identified categories and the small sample from the first step, the categorizations by research assistants and the categorizations by MTurk workers demonstrated high reliability. The seven categories seemed to adequately reflect UGC content characteristics—at least in social media business pages—with a large proportion being important from a management perspective (Yang et al. 2019).

3 Collecting UGC

In recent articles, Boegershausen et al. (2022) as well as Guyt et al. (2024) discuss the opportunities and challenges associated with collecting web data for marketing purposes. In particular, they emphasize the richness of this data source from a content perspective, the diversity of the available data types, and the integration problems with traditional data. Especially, they highlight that web crawlers, web scrapers, and APIs (application programming interfaces) are the most important techniques, but that these approaches require both technical expertise and substantial resources. Additionally, they stress that overcoming technical barriers related to data extraction, storage, and processing demands further creativity, especially given the typically large volume of data involved. In the following subsections, we discuss these three main techniques for collecting UGC as described in the preceding section.

3.1 Web Crawlers and Web Scrapers

Simply put, web crawling is the automated process of systematically navigating and browsing the web to discover and index web pages by following the links between and to them. In contrast, web scraping involves extracting specific pieces of data—UGC, in the present context—from web pages or social media by parsing their content and retrieving the desired information (Khder 2021). Web crawlers are designed to handle the dynamic nature of web content, accessing both visible and hidden data, and they often incorporate techniques for managing vast amounts of information. Web scrapers, on the other hand, are essential for data-driven decision-making because they may provide accurate and efficient solutions for data extraction across various web environments. In marketing, both techniques can be used—for example, to track brand mentions across the web, gather competitive intelligence, monitor market trends, or analyze consumer sentiments to make informed decisions and optimize marketing strategies (Levene and Poulovassilis 2001). Table 1 provides a concise overview of web crawlers discussed in the relevant literature (see, e.g., Amuhda 2017; Deshmukh and Vishwakarma 2021; Gupta and Anand 2015; Menczer et al. 2004; Thenmalar and Geetha 2011; Dhenakaran and Sambanthan 2011; Bergman and Popov 2023). A common feature of all these web crawlers is their objective to capture the largest possible amount of data—in this case, UGC—in the shortest possible time.

Similarly, Table 2 provides a concise overview of web scrapers discussed in the relevant literature (see, e.g., Darmawan et al. 2022; Dellarocas et al. 2013; Egger et al. 2022; Sharkey et al. 2023) and on relevant websites (e.g., www.brightdata.com, www.techopedia.com). Multiple implementations exist for most of these scrapers, some developed by the scientific community and others offered by commercial organizations.

Guyt et al. (2024) additionally distinguish between code based web scrapers, non-code based scrapers and LLM (large language model) based scrapers according to the tools applied:

- Code based web scrapers are programmed using popular and free of cost programming languages/environments like R (www.r-project.org) or Python (www.python.org). These integrated development environments offer, besides their basic functionality for programming, statistical computing, and visualization, many extension packages. So, for web scraping, R offers the packages rvest for HTML parsing and Selenium for headless browser scraping. Python offers the packages BeautifulSoup and Scrapy for HTML parsing and selenium for headless browser scraping.
- Non-code web scrapers like Octoparse (www.octoparse.com), Import.io (www.import.io), ParseHub (www.parsehub.com), and WebHarvy (www.webharvy.com) are commercial tools which provide an easier data collection access to UGC than code based web scrapers. A visual interface is provided that simulates the usual access to websites including clicking and typing of text. By marketing example website elements during this supervised access, the user identifies the desired information, starts the extraction, and receives the scraped web data in short time, formatted, e.g., as Excel files. Using non-code web scrapers typically comes at costs, depending on the data volume collected.

Table 1 Overview of web crawlers for UGC data collection

Name	Description
Generic web crawlers	Generic web crawlers aim to search the web and index accessible web pages without any specific focus. They are primarily used by search engines but are unable to crawl the vast amounts of data presented in the hidden web.
Focused web crawlers	These crawlers are specialized in retrieving web pages related to specific topics, thereby reducing the effort spent on viewing websites that are unlikely to provide relevant information. However, they often fail to target the content of the hidden web.
Incremental web crawlers	Incremental crawlers are designed to keep their data updated by revisiting pages and refreshing them only when changes have occurred since the last crawl. They focus on updating already indexed content rather than recrawling the entire web.
Distributed web crawlers	This type of crawler operates across multiple machines or nodes, coordinating to maximize efficiency and balance the load. The parallelization of the crawling process makes these crawlers particularly effective for large-scale web crawling operations.
Parallel web crawlers	These crawlers run multiple processes simultaneously to cover more ground in less time. This approach is essential when dealing with massive datasets, that is to say large numbers of web pages, and when time is a critical factor.
Ontology-based web crawlers	These crawlers use domain-specific ontologies to guide the crawling process. They help to retrieve more precise data by estimating the semantic content of a URL link in a set of documents based on the domain-dependent ontology.
Path-ascending crawlers	These crawlers ascend the directory structure of a URL, moving from specific files to broader directories. They are particularly effective in discovering isolated resources or those without inbound links.
Shark-search crawler	Shark-search crawlers are an enhanced version of the fish-search algorithm, designed to discover relevant information more efficiently. They use a specific scoring system to evaluate and rank links/documents based on their relevance.
Dark/hidden web crawlers	Crawling the dark/hidden web to extract data from hidden services is a complex process that requires specialized methodologies and techniques. This type of web crawler can be used, for example, to investigate and anticipate potential cyber threats.

- LLM based scrapers make use of recent advances in Transformer-based deep learning (see Sect. 4 for details): Chatbots like ChatGPT-3.5 or ChatGPT-4o (OpenAI 2024) can be instructed in natural language to collect UGC data for a specified app, product, or service, or to extract UGC from a specified web page. So, e.g., if we know the web page where the retailer Otto offers the fridge-freezer combination CNsdc 5203_994876651 (www.otto.de/p/liebherr-kuehl-gefrierkombination-cnsdc-5203_994876651-185-5-cm-hoch-59-7-cm-breit-1679410244/#ech=28954569&variationId=1716996242) and we know which web page contains corresponding OCRs (www.otto.de/kundenbewertungen/1679410244/#variationId=1716996242, accessed by clicking on the element "reviews"), we can ask ChatGPT to download all reviews from this web page into an Excel file with the following prompt: "Please collect all reviews from the web page www.otto.de/kundenbewertungen/1679410244/#variationId=1716996242 and store them in the Excel file ocr.xlsx." It should be mentioned that, currently, this chatbot access is sometimes error-some (e.g., collects only subsamples of OCRs on addressed web pages) and is—by far—not as flexible as the access by code or non-code based tools discussed above (looping across websites and web pages is incomplete). Here, one could ask, alternatively, ChatGPT or other LLM based chatbots to develop an R or Python code for this purpose and execute this programming code. In a similar way, LLM based chatbots can be used as content aggregators to summarize OCRs on apps, products, or services (see Subsect. 4.3 on this issue). However, despite their potential benefits, most AI-powered review summarization tools have been evaluated primarily in experimental settings using datasets such as Amazon or Yelp reviews, whereas their use for data collection in real-world settings is still in its infancy.

Table 2 Overview of web scrapers for UGC data collection

Name	Description
HTML parsers	HTML parsers process HTML code by analyzing it and converting it into a structured data format, known as the Document Object Model (DOM). This format allows specific elements to be extracted, such as tags, headings, and paragraphs.
Headless browser scrapers	Headless browser scrapers simulate a full browser environment without a graphical user interface running in the background. They can interact with web pages as a user would, including executing JavaScript and handling dynamic content.
Text-based scrapers	Text-based scrapers focus on extracting raw text content from web pages, often using regular expressions (regex) to find and extract specific information. These scrapers are useful for straightforward, text-heavy, web pages.
AI-powered scrapers	These scrapers employ machine learning techniques to identify and extract relevant information based on learned patterns. They can adapt dynamically to web page structures, making them effective for scraping complex and changing data.
Content aggregators	Content aggregators compile data from multiple web sources into a unified dataset, commonly used for tasks such as collecting product prices across e-commerce sites and aggregating OCRs from various platforms.

3.2 API-Based Methods

An often somewhat neglected alternative for extracting UGC, or marketing data in general, from the Internet is the use of APIs (Boegershausen et al. 2022). Simply put, an API allows different software systems to communicate by sending requests and receiving responses (www.ibm.com/topics/api). It acts as a bridge between a client (which sends the request) and a server (which provides the data, such as UGC). APIs enable seamless integration, allowing developers to utilize external functionalities without needing to build them from scratch, while keeping internal systems secure and hidden.

APIs offer several advantages in the present research context, but they also come with disadvantages (see, e.g., Boegershausen et al. 2022; Maleshkova et al. 2010; Lomborg and Bechmann 2014; Ruelens 2022; Puschmann and Ausserhofer 2017). Advantages are:

- They provide an authorized, standardized, and structured way to access data (in real-time), particularly when a data provider explicitly offers specific APIs for programmatic access to its database(s). Examples include the Yelp API (www.docs.developer.yelp.com) and the Holiday API (www.holidayapi.com).
- They often offer a legally and ethically acceptable method for collecting UGC data, helping researchers avoid issues related to data scraping or unauthorized access to UGC, as their use is typically covered by the terms of use of the platforms providing the data.
- APIs typically support various input parameters and output formats (such as XML and JSON), making them adaptable to different data needs and systems.
- The widespread availability of APIs across various domains, including social media, allows for extensive data access, particularly from popular platforms, such as Facebook, Yelp, and X.
- APIs facilitate the automation of UGC data collection processes and the reuse of code, enabling more efficient and scalable data gathering.

Disadvantages are:

- The absence of well-established standards in API documentation and usage may result in inconsistent practices, occasionally requiring significant customization.
- UGC data collected via APIs also (as the discussed web crawlers and web scrapers) may suffer from non-representative sampling, potentially affecting the validity of research findings.
- Companies may restrict API access for strategic reasons, limiting the availability of certain UGC data types, thereby impacting the comprehensiveness of data collection.
- Even when using APIs, researchers may encounter ethical challenges, particularly concerning privacy, because the providers of UGC often do not explicitly consent to their data being used in research, even if the data are publicly accessible.
- The use of APIs may come along fees for usage, especially for high volumes of data to be collected.

For the sake of completeness, it should be mentioned that collecting UGC data without explicit user consent may raise significant privacy issues, especially when personal data is involved, bringing data protection regulations such as GDPR (General Data Protection Regulation) and CCPA (California Consumer Privacy Act) into play, or when other complex ownership rights apply (Xiao 2020).

As a summary to the opportunities and challenges of collecting UGC data, Boegershausen et al. (2022) emphasize that it is crucial to consider technical, legal, and ethical issues to ensure the validity and utility of collected UGC data for advancing marketing insights. The authors provide a methodological framework for collecting web data, which essentially comprises three steps: (1) source selection, (2) collection design, and (3) data extraction. Boegershausen et al. (2022, p. 5) contend that "these decisions often involve trade-offs about research validity, technical feasibility, and legal/ethical risks that are not always apparent".

4 Analyzing UGC

4.1 Data Analysis and Machine Learning Based On UGC Feature Extraction

4.1.1 UGC Feature Extraction

In contrast to many other data sources that can be analyzed to support marketing decisions, UGC (e.g., web-scraped from blogs, online shops, or social media networks) is often characterized as unstructured for two reasons (Balducci and Marinova 2018; Wedel and Kannan 2016):

- UGC single data units are often non-numeric (e.g., OCRs that consist of a comment and/or audio, media platform posts on media hosting sites collecting photos and/or videos) and must be pre-processed (transformed to meaningful numerical values and/or binary indicators) to be analyzable.
- UGC single data units contain multiple facets/aspects simultaneously (photos and/or videos can be compared through, for example, their color/edge distributions and/or the number of contained faces; audio through, for example, its pitch, speed rate, and/or speech intensity), which means that an analyst must decide on which facet/aspect to focus during an analysis.

The non-numeric and multi-faceted characteristics of UGC prevent traditional data analysis and machine learning (e.g., regression analysis, decision trees, and neural networks) from being directly applied, as is common with structured data (e.g., data matrices with meaningful numeric values and/or binary indicators). However, for a long time, feature extraction has been a successful solution for these pre-processing requirements. Thus, Fayyad et al. (1996) discuss the finding of useful features that describe single data units as an important part of the knowledge discovery in databases (KDD) process before methods can be applied.

For natural language texts or transcribed audio as single data units (Feldman and Dagan 1995; Tabassum and Patil 2020), textual pre-processing has proven to be useful (e.g., change to lower cases, stop words removal, removal of punctua-

tions, removal of numbers, stemming), followed by one or more widespread feature extraction techniques.

- Term/named entity recognition: The single data unit is scanned for terms pre-specified by words or *n*-grams (combinations of words). Binary or count features for these terms indicate their occurrence in the text.
- Bag-of-words coding: The occurrence of frequent words or *n*-grams in a single data unit is coded by binary or count features for each word or *n*-gram.
- Term frequency-inverse document frequencies (TF-IDF) of words or *n*-grams coding: TD-IDF coding is similar to bag-of-words coding, but the relative frequency of a word or *n*-gram in the single data unit is multiplied by the logarithm of the inverse relative frequency of all single data units containing this word or *n*-gram. The second factor reflects how much information the word or *n*-gram provides, i.e., how common or rare it is across all single data units.

For images—coded as matrices of intensities per pixel (e.g., one value for the grey intensity per pixel or three values in the red-green-blue color model per pixel)—other extractable features have long been used (for an overview see, e.g., Baier et al. 2012).

- Color distributions: One- or three-dimensional histograms reflect the frequency of intensity ranges.
- Edge, texture, or shape distributions: Features indicate the frequency of specific edges, textures, or shapes in specified parts of the image.
- Named entity recognition: The image is scanned for pre-specified named entities (e.g., objects or persons). Binary or count features indicate their occurrences.

The main advantage of the UGC feature extraction approach (both for texts and images) is the possibility that the resulting data matrices (with individuals or UGC single data units as rows and features as columns, containing numeric or binary values) now enable the application of traditional data analysis and machine learning methods.

4.1.2 Traditional Data Analysis and Machine Learning Based On UGC Features

After the feature extraction process, the resulting numeric data matrix can be analyzed in a traditional manner. For this purpose, the following data analysis and machine learning methods for unsupervised and supervised learning are widespread (Wedel and Kannan 2016; Ma and Sun 2020; Ngai and Wu 2022; Herhausen et al. 2024; Duarte et al. 2022).

- Unsupervised learning: e.g., term frequency analysis/word clouds, k-means, neural networks like self-organizing maps (SOM), topic modeling using latent Dirichlet allocation (LDA).
- Supervised learning: e.g., lexicon-based approaches, (negative binomial, logistic, probit, or tobit) regression analysis, k-nearest neighbor, naïve Bayes analysis, decision trees/random forests/XGBoost, support vector machines, neural networks like multi-layer perceptrons.

Term frequency analysis based on term recognition features is an often applied unsupervised learning method when the intention is to identify frequently discussed aspects, strengths, or weaknesses. For example, Kim et al. (2022) used this method to derive three lists of frequently occurring terms in OCRs: one list across all OCRs, a second list across OCRs with 5-star ratings, and a third list across OCRs with 1- or 2-star ratings. The comparison of these lists helped to identify intensively discussed aspects as well as app, product, or service attributes that lead to high or low satisfaction. Word clouds (visualizations of term frequencies) supported these findings.

Topic modeling can also be applied to OCRs for this purpose: Topics are defined as a probability distribution across words (or *n*-grams) of a vocabulary. For example, if some customers use in their OCR comments words like "clean", "quiet", "spacious", "soft" instead of words like "meal", "breakfast", "sightseeing", one could assume that these customers reflect the topic "hotel room quality" and not topics like "food quality" or "location". For estimating topics and corresponding customer segments from a sample of OCR comments (or other texts), LDA, a statistical model introduced by Blei et al. (2003), assumes that the probability that a word (or *n*-gram) occurs in an OCR depends on three types of probabilities: The probabilities that the customer belongs to specific customer segments, the probabilities that these customer segments discuss specific topics, and the probabilities that the word is used in an OCR comment when these topics are discussed. The underlying probabilities (model parameters) can be estimated by applying maximum likelihood or other statistical approaches to the OCR comments and allow to allocate—in a probabilistic manner—customers to (up to now unknown) customer segments, words to (up to now unknown) topics, and topics to customer segments.

When the data matrices contain extracted features from images, allocating individuals to customer segments is also possible: So, e.g., Baier et al. (2012) and Daniel (2014) used k-means to cluster individuals based on similarities of their uploaded photos during a survey. Each individual and her or his uploaded images were represented by category-specific extracted color diagrams (reflecting her or his holiday/spare time, apartment/furniture, and clothing/fashion interests). The authors found that an additionally conducted traditional lifestyle survey among the same individuals, with multi-item scales for activities, interests, and opinions, resulted in similar lifestyle segments.

Supervised learning methods, in contrast to the up to now discussed unsupervised ones, often train a model that predicts an interesting outcome based on features of the data matrix. The outcome could be the overall positive or negative sentiment of an OCR or the indication whether a photo contains a brand logo or not. For a small sample of single data units, the corresponding outcomes are known. Then, a model is trained and tested to predict the outcome based on corresponding extracted features. Finally, the trained model is applied to all data units. Some of the earliest approaches to predict sentiment scores from texts was lexicon-based sentiment analysis, So, e.g., Pennebaker et al. (2001) introduced the text analysis software LIWC in the early 2000s for studying the various emotional, cognitive, structural, and process components present in text samples. LIWC relied on a lexicon with 4500 words which were assigned to 76 categories, including 406 words that indicate positive emotion

(e.g., love, nice, good, great) and 499 words that indicate negative emotion (e.g., sad, bad, worse). The difference between the relative frequencies of these indicators in an OCR comment is then used to predict an overall sentiment score. More advanced lexicon-based sentiment score predictors like VADER (Hutto and Gilbert 2014) allocate polarity values to the words in their lexicon. The score prediction then relies on calculated sums of the polarity values of occurring words in a comment. Such lexicons, if properly built, allow to be used as sentiment predictors across a wide range of textual UGC. They have remained popular even after the introduction of machine learning based alternatives (Humphreys and Wang 2018). However, it has to be mentioned that lexicon construction is often time intensive, domain specific, and prone to a wide variety of difficulties (Islam and Zibran 2018).

Besides these lexicon-based approaches, as already mentioned, a large number of traditional machine learning methods is available for predicting sentiment scores or other outcomes (see the overviews in Wedel and Kannan 2016; Ma and Sun 2020; Ngai and Wu 2022; Herhausen et al. 2024; Duarte et al. 2022): Decision trees are built by iteratively splitting the training sample of single data units according to feature values so that homogeneous subsamples are formed with similar outcome values. (Artificial) Neural networks—e.g., multilayer perceptrons—were inspired by the structure and function of the human brain with its neurons (as nodes) and connecting synapses (as weights): Feature values at nodes of a so-called input layer are weighted to give values at nodes in hidden layers. Then, these values are weighted to give outcome values at nodes in the output layer. The training sample with known corresponding feature and outcome values is used to estimate the weights that connect the nodes of neighbored layers so that the predicted and the observed outcomes match as best as possible.

Other methods, e.g., random forest or XGBoost, combine simpler models (e.g., decision trees) to stabilize their prediction. For a long time and even today, these traditional machine learning methods based on extracted features have repeatedly demonstrated their predictive accuracy. For example, Salminen et al. (2022) collected in their comparison of machine learning methods for UGC data analysis $n =$ 4,209,101 tweets on 20 brands. They extracted TF-IDF features for each tweet in order to predict brand-specific pain points (important customer concerns) from these tweets. Two researchers were asked to independently label a train and test sample of 2000 tweets with binary outcome values that indicated whether the tweet contained a pain point or not. Then, various supervised learning methods were applied to train and test a pain point prediction model. Here, the naïve Bayes method performed best, followed by XGBoost and the k-nearest neighbor method. The accuracy of the naïve Bayes method was only slightly outperformed by an additionally applied modern discriminative deep learning method using the Transformer architecture (discussed in the next subsection). Finally, the prediction model was applied to all tweets. Tweets with highest pain point predictions then were used to develop improvement hints for the analyzed brands.

Besides machine learning methods to train and test a prediction model from UGC as discussed, wide-spread are also applications of regression analysis for the same purpose. For example, Decker and Trusov (2010) analyzed mobile phone OCRs by extracting 46 attribute-level binary indicators as terms (i.e., whether 23 selected

attributes like size/weight, appearance, functionality, and battery were positively or negatively discussed in a comment) and related them to stated preference (here: the overall star rating in the OCR) by applying a negative binomial regression model. The impact of the attribute-level indicators on preference could then be used to understand the importance of attributes in customers' mobile phone evaluations. Other authors applied other regression variants to similar data. For example, Kübler et al. (2024) analyzed the helpfulness of OCRs (rated by other users) using the so-called tobit regression. Here, due to the conditional nature of the helpfulness rating (users can only give positive helpfulness indications and rate only few OCRs), a censored form of regression—the so-called tobit regression—had to be applied.

However, despite its widespread use and accuracy, the feature extraction approach discussed so far has been widely criticized because it often focuses on wrong facets/aspects in the UGC single data units (specific text or image features) and leads to misestimations of single data unit (dis)similarity. For example, Baier et al. (2012) discuss the problem that images have similar color/texture distributions but different content (e.g., a photo with the upper body of a young woman with curly hair and a photo with a cocker spaniel had a similar color/texture distribution). Similarly, the bag-of-words approach can lead to misjudgments—for example, when the same words are used in an ironic context, when negations are ignored ("not easy," "no cloud storage"), or when the sequence of words makes a difference. An alternative here is to use advanced machine learning methods that omit the feature extraction step, as described in the following subsections.

4.2 Discriminative Deep Learning

4.2.1 Image-Based Discriminative Deep Learning Using Convolutional Neural Networks

Since the 2010s, neural networks with large numbers of layers, nodes, and—consequently—parameters (weights, scaling parameters), so-called "deep" neural networks, dominate in machine learning challenges, when unstructured data (e.g., text, image, audio, and video) has to be classified (see, e.g., Chollet 2021, p. 16ff.). In 2012, the ImageNet challenge to recognize 1000 predefined object categories (e.g., balloon, strawberry) in 1.4 M images was won by Alex Krizhevsky and Geoffrey Hinton. Their convolutional neural network (CNN) AlexNet had eight layers, 650,000 nodes, and 60 M parameters, and enabled a breakthrough accuracy of 83.6% in this challenge (Krizhevsky et al. 2017). Since then, most similar challenges with a discriminative task (natural language text or image classification) have been won by deep neural networks. CNNs, the typical winners in image classification tasks in the 2010s, consist of an input layer, subsequently followed by multiple hidden layers, and finally an output layer, all with large numbers of nodes. It has to be noted that at the input layer for each pixel one input node (if grey intensities are the input) or three (if red-green-blue or other color model intensities are the input) are needed. The CNN self-extracts feature values from these input nodes by applying standardized filter operations to small parts of the input data in the input layer (e.g., all 3×3 neighbored pixel areas of an image or other neighbored tokens) and by

using the outputs of a layer's nodes as inputs in a similar fashion to the nodes of the next layer. Finally, in the output layer, for each category to be predicted, a node calculates a value based on the values at the nodes of the previous layer.

Since such deep neural networks consist of huge numbers of unknown parameters (weights, scaling parameters), enormous amounts of train/test data, computation time, and storage space are needed. However, from the beginning, it was seen to be meaningful to fine-tune already trained deep neural nets to new categorization tasks since this transfer of an already estimated model reduces data, time, and storage requirements significantly. CNNs trained to predict the 1000 categories of the ImageNet challenge could be used as a basis to predict other (similar) categories by replacing the trained output layer with a new output layer—to be trained independently and fine-tuned—that predicts the categories of the new task. A well-known, ImageNet pre-trained CNN model that is available in R and Python is VGG16 (Simonyan and Zisserman 2015). This pre-trained model has been fine-tuned for many other contexts—for instance, to predict brand confusion in imagery markets from print ads and from TV video ads (Nakayama and Baier 2020) or to detect brand logos in posted UGC (Wang et al. 2022). The fine-tuning of pre-trained models for many contexts is called the application of transfer models or transfer learning (Weiss et al. 2016).

4.2.2 Text-Based Discriminative Deep Learning Using Recurrent Neural Networks

Whereas CNNs dominated the image-based machine learning challenges in the 2010s (see, e.g., Chollet 2021, p. 16ff.), deep learning based on so-called recurrent neural networks (RNNs) did the same for text-based tasks like speech recognition, machine translation, and text categorization (see Sutskever et al. 2015). Unlike multilayer perceptrons and CNNs that follow a feedforward mechanism and process data in a single pass, RNNs process data across multiple sequence or time steps, making them well-adapted for modeling and processing sequential or time-series data like text, audio, and/or video (Tealab 2018; Sarker 2021). This modeling endeavor is achieved by employing nodes with a hidden state, essentially a form of

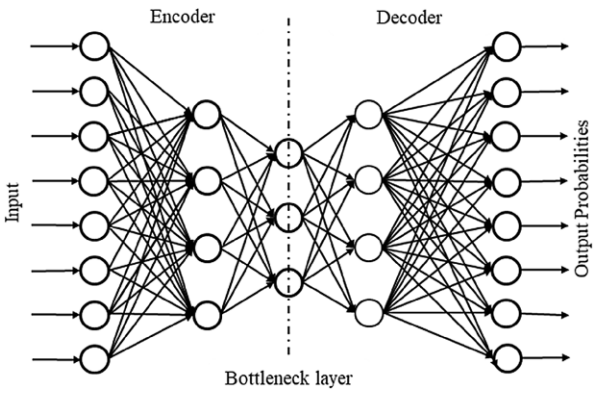

Fig. 1 Encoder-decoder principle in machine learning, non-recurrent version. (Adapted from Rumelhart et al. 1987)

memory, which is updated at each sequence or time step based on the current input and the previous hidden state (Hochreiter and Schmidhuber 1997). This recurrency allows the network to learn from past inputs but increases substantially the time the model needs to be trained. The well-known so-called long short-term memory (LSTM) RNN variant (Hochreiter and Schmidhuber 1997) became the default architecture for natural language processing with—compared to other RNNs—acceptable training time and high accuracy in text classification tasks.

Newer so-called deep RNNs with multiple LSTM layers (see, e.g., Sutskever et al. 2015; Sachin et al. 2020; Yadav et al. 2023) demonstrate even better accuracy for text classification than RNNs with few LSTM layers. They are based on the well-known encoder–decoder principle (Ackley et al. 1985; Rumelhart et al. 1987), visualized in Fig. 1: the network consists of successive layers of nodes that are interconnected. The encoder part of the network transforms high-dimensional data from input nodes to low-dimensional representations in bottleneck layer nodes. The decoder part 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 using large samples of given input and output data pairs so that the calculated outputs from the given input data are as close as possible to the corresponding given output data.

In its simplest (non-recurrent) form as in Fig. 1, the encoder-decoder principle is trained with given pairs of identical input and output data to estimate the relations between high-dimensional representations and (unknown) low-dimensional ones without a 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- 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, this encoder-decoder principle has established itself as the superior architecture for RNNs, especially when text (as a sequence of words or word pieces) has to be translated from one language to another (see, e.g., Bahdanau et al. 2014; Sutskever et al. 2015; Cho et al. 2014). Based on this principle, so-called large language models (LLMs) were trained, which then can be fine-tuned—again called transfer models as with CNNs—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).

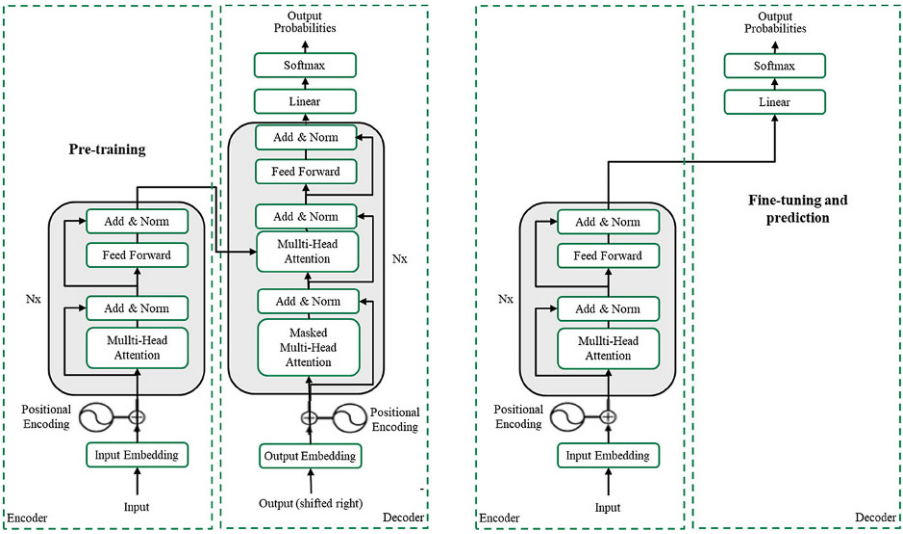

Fig. 2 Illustration of the main components of the Transformer architecture (*on the left side*, adapted from Vaswani et al. 2017) and its application as a transfer model to text categorization (*on the right side*)

4.2.3 Text-Based Discriminative Deep Learning Using the Transformer Architecture

Since then, however, the Transformer architecture—first presented in 2017 (Vaswani et al. 2017)—has further developed the accuracies and the affordable tasks of machine learning being nowadays—instead of CNNs or RNNs—the gold standard for developing transfer models (Chang et al. 2024). LLMs trained in this architecture differ from CNNs and RNNs insofar as they completely abstain from convolution and recurrency. Instead, they solely rely on the so-called attention mechanism as a feedforward concept that generates LLMs "to be superior in quality while being more parallelizable and requiring significantly less time to train" (Vaswani et al. 2017, p. 1). The attention mechanism allows to focus on the most relevant parts of the input by assigning varying importance to different elements. It calculates attention scores and applies a softmax function to create probabilities that weight the value matrix. This enables the model to capture relationships across the entire sequence efficiently, even over long distances.

The left side in Fig. 2 (adapted from Vaswani et al. 2017) illustrates the main components of this Transformer architecture applied to natural language tasks:

- Input and output data are texts (a sequence of words) that are converted by so-called tokenizers according to a vocabulary (list of frequent and meaningful words or word pieces in languages) into indicators of tokens (words or word pieces).
- The tokens and their positioning (sequence information) are then converted to 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 Transformer layers (e.g., Nx = 6) carry out repeated transformations on these vector representations to extract more and more abstract linguistic information. Each Transformer layer consists of an attention layer and a feedforward layer. These attention layers are specific to the Transformer architecture. They are specific matrix operations that enable the learning of token or node amplifications depending on the context as proposed by Bahdanau et al. (2014). However, in contrast to former attention propositions with RNNs, the Transformer architecture achieves this in a feedforward manner and, therefore, requires less training time.
- An optional, un-embedding layer converts the final vector representations back to a probability distribution of the tokens.

To pre-train Transformer-based LLMs, large text corpora were used as a basis for input and output data pairs. The 2014 Workshop of Machine Translation (WML) dataset with about 4.5 million English-German sentence pairs should be mentioned (Vaswani et al. 2017) as should the Toronto BookCorpus 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), and a (machine) translation task (with the same text in two different languages as input or output text pairs) (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. Vaswani et al.'s (2017) basic LLM ("base") makes use of a word-piece vocabulary with 25,000 tokens, allowing up to 512 tokens as input and output text (text with up to 512 word pieces). Nx = 6 Transformer layers in the encoder and in the decoder parts produce 65 million network parameters to be estimated. The largest LLM ("big") with 1024 tokens as input and output is determined as having 213 million network parameters to be trained (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 have even higher demands on memory volume and computing time.

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 were trained originally (e.g., machine translation, language inference, and question answering). Thus, the nowadays ubiquitous BERT (Bidirectional Encoder Representations from Transformers; see Devlin et al. 2018) is a widespread family of LLMs that mainly consist of the encoder part 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 datasets—is now used, for the most part, as a transfer model for text categorization (e.g., for sentiment analysis). The right side in Fig. 2 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 but are then directly led by simple transformations to output probabilities for text categories (e.g., input is "acceptable"

or "not acceptable"). Like many other LLMs, a BERT model was trained by using the Toronto BookCorpus with 800 million words and the English Wikipedia corpus with 2500 million words. However, instead of usually training BERT with pairs of input and output data according to the encoder-decoder principle, BERT receives for training corresponding (and, alternatively, not corresponding) input/output text pairs as single input data. The transformation parameters are learned by comparing the output probabilities for the categories "acceptable" and "not acceptable" with the information on whether the input consisted of a corresponding and not corresponding input/output text pair. This modification of the basic Transformer architecture has proven to be advantageous due to its simultaneous/bidirectional analysis principle, especially when text categorization is the major task for which the LLM is to be applied (for details see Devlin et al. 2018). Recently, BERT has been extended to allow sentiment analysis with multilingual texts (especially applicable for texts in Dutch, English, French, German, Italian, and Spanish). 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 within 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.

4.2.4 Image-Based Discriminative Deep Learning Using the Transformer Architecture

In the beginning, Transformer-based LLMs were solely applied as transfer models to solve natural language tasks. However, recently, they have demonstrated their superiority over other deep learning approaches in the field of image and video classification (Dosovitskiy et al. 2021) as well as in image and video generation (Parmar et al. 2018). Dosovitskiy et al. (2021) proposed the Transformer-based VisionTransformer (ViT) that, compared to well-known pre-trained and fine-tuned CNNs in the context of well-known image classification tasks, needs less training time and provides superior performance (accuracy rates in image classification). For example, Kim and Moon (2024) compared ViT with a state-of-the-art CNN (a VGG16 network, pre-trained based on the ImageNet database) with respect to their accuracy when classifying images with bulky waste. They found that ViT clearly outperforms the CNN (83.89% accuracy versus 64.60% accuracy). Truong and Lauw (2023) compared ViT and two well-known CNNs (ResNet-152 and EfficientNet-B7) for sentiment score prediction based on OCR images and found a similar superiority of ViT.

Google's recently introduced Transformer-based Gemini family of highly capable multimodal models (Gemini Team 2024) goes even one step further by allowing to understand and analyze combinations of text, image, audio, and video. Pre-trained with data from web documents and articles including image, audio and video, the

models can be fine-tuned to analyze multimodal data for specific tasks like classification of multimodal data units, question-answering, reasoning, or image and text generation. The comparisons showed that these Transformer-based models outperform all other discriminative deep learning models (Gemini Team 2024).

4.3 Generative Deep Learning

Besides discriminative tasks like text, image, audio, or video classification, Transformer-based LLMs and derived chatbots have demonstrated additional possible uses (Chang et al. 2024): Due to their trained ability to iteratively predict adequate next words based on current context, they are able to interact with humans in a convincing manner, giving human-like answers to questions or fulfilling other tasks like text summarization or drawing conclusions from presented data: The human provides a prompt in natural language that contains a question or instruction combined with data and a desired output format. The LLM or chatbot responds in short time and with impressive results. So, in their comprehensive overview of the performance of current Transformer-based LLMs and chatbots, Chang et al. (2024) found that LLMs like GPT-3.5 or GPT-4 (GPT stands for Generative Pre-Trained Transformer, OpenAI 2023), as well as derived chatbots like ChatGPT-3.5 or ChatGPT-4o (OpenAI 2024), are convincing—even in comparison to human experts—when it comes to answering questions or instructions fluently and correctly in terms of content and language, summarizing and classifying presented collections of texts and images, and drawing desired conclusions. However, Chang et al. (2024) also found, that up-to-now LLMs and derived chatbots are still inferior to humans when it comes to capturing inconsistencies and semantic subtleties in texts and images. They are susceptible to misinterpretation and, if calibrated further back, are often not at the current level of information and knowledge required to correctly answer a question or instruction. Recent applications in a real-world business context (e.g., Bouschery et al. 2023, Jeong and Jihwan 2024) confirmed these findings: Bouschery et al. (2023) found that GPT-3.5 was able to summarize OCRs adequately and to develop ideas and concepts for product/improvement based on them, Jeong and Jihwan (2024) demonstrated that ChatGPT-3.5 was able to answer adequately to customers' complaints in TripAdvisor OCRs in a hotel service context.

Given the ever-increasing volume of information available on the Internet regarding consumer behavior, opinions, sentiments, recommendations, and so on, generative deep learning is becoming an increasingly important method for aggregating and analyzing UGC. One particular type of aggregation that is currently receiving significant attention is OCR summarization: Large numbers of OCRs are condensed to concise summaries that capture key sentiments and opinions. This helps consumers make quick decisions and provides businesses with actionable insights (Jovanovic and Campbell 2022; Dwivedi et al. 2023). Frequently used techniques in this context include (Widyassari et al. 2022; Zhao et al. 2022; George and Srividhya 2022; Uppalapati et al. 2023; Zhang et al. 2021b; Xu et al. 2023; Wang et al. 2020), e.g.

- extractive summarization: Selecting and combining key sentences or phrases from OCRs,

- abstractive summarization: Generating new sentences that capture the essence and key points of OCRs, as well as
- hybrid approaches: Combining extractive and abstractive methods to leverage the strengths of both approaches.

These OCR- or UGC-based summaries offer several benefits for decision support in corporate marketing. They enable marketers to quickly understand customer sentiments and key concerns, helping marketing teams make informed decisions by highlighting common feedback themes. This, in turn, can lead to better product improvements and more effective marketing strategies. Additionally, automated summarization saves time by efficiently processing vast amounts of review data, and these techniques are scalable across languages and regions, making them invaluable for global marketing. In conclusion, review summarization is likely to play a vital role in future marketing decision support by providing actionable insights and improving decision making.

In the next section, we provide a short application overview of the traditional and the advanced methods discussed to support decisions in marketing management. We discuss successful use cases for UGC data collection and analysis for different marketing management decision areas based on the approaches discussed (Sect. 5), and we summarize the findings in a guideline for UGC data collection and analysis (Sect. 6).

5 Use Cases of UGC-Based Decision Support in Marketing Management

5.1 Text-Based Use Cases

The analysis of text-based UGC has a long tradition in marketing management (Balasubramanian and Mahajan 2001; Decker and Trusov 2010; Timoshenko and Hauser 2019). Table 3 gives an overview of decision areas, pursued goals, data sources, and applied methods. Among the huge number of articles that deal with UGC data analysis (Li et al. (2022) refer in their literature review to more than 3390 published academic research articles for UGC data analysis in e-commerce), we focus on articles where—in contrast to discussing the methods applied—the focus is on marketing decisions. Additionally, we discuss here shortly (as already done in Sect. 4 for some other use cases from Table 3 refering to the methods applied, e.g., Decker and Trusov 2010; Jeong and Jihwan 2024; Kim et al. 2022; Kübler et al. 2024; Salminen et al. 2022) some selected use cases.

A closer look at the use cases in Table 3 shows that text-based UGC use cases are often in decision areas like customer/market selection, product/service quality management, new product/service development, customer relationship management, and brand management. By applying traditional data analysis and machine learning as well as discriminative and generative deep learning to collected UGC data, it is possible to identify important product attributes and topics from a customer's perspective.

Table 3 Overview of text-based use cases of decision support in marketing management

Decision area	Pursued goal	Collected UGC data	Applied method(s)	Reference
Customer/market selection	Identify important cell phone attributes for segments	$n = 20{,}419$ cell phone OCRs with pro/con summaries and rating	Extraction of attributes and their valence from comments; application of negative binomial regression analysis	Decker and Trusov (2010)
Product/service quality management	Improve acceptance of an online shop app (IKEA)	$n = 755$ online shop OCRs with comments and rating	Prediction of perceived usefulness, ease of use, and other TAM constructs using a lexicon-based approach	Rese et al. (2014)
Product/service quality management	Identify relevant topics for hotel guest satisfaction	$n = 696$ restaurant, $n = 4467$ hotel OCRs (text, rating)	Topic modeling using LDA based on comment parts (sentences)	Büschken and Allenby (2016)
Customer/market selection	Identify important cell phone attributes for segments	$n = 679{,}422$ cell phone OCRs (text, rating)	Extraction of attributes and their valence using a lexicon-based approach; prediction of helpfulness/importance	Qi et al. (2016)
Customer/market selection	Identify important cell phone attributes for segments	$n = 2245$ cell phone OCRs with pro/con summaries and rating	Extraction of attributes and their valence using a lexicon-based approach; prediction of attribute importance	Xiao et al. (2016)
Product/service quality management	Identify important attributes (e.g., battery, service)	$n = 3845$ laptop and $n = 3841$ restaurant OCRs	Extraction of targets (e.g., battery, service) and the corresponding opinion using recurrent neural networks	Liu et al. (2018)
Product/service quality management	Predict user ratings from OCRs to monitor satisfaction	$n = 1557$ open source software OCRs	Sentiment analysis of comments using a pre-trained GloVe word embedding (using Amazon reviews) and LSTM	Gezici et al. (2019)
New product/service development	Identify consumer needs for new oral-care products	$n = 115{,}099$ oral-care OCRs (text, rating)	Training a CNN to predict informativeness based on comments coded by experts; application to all comments	Timoshenko and Hauser (2019)
Brand management	Monitor brand awareness, impression, purchase intention, satisfaction	$n = 27{,}956$ brand-day-observations with Facebook comments on 48 brands	Sentiment analysis of comments; train/test vector autoregressive models to predict brand awareness, impression, purchase intention, satisfaction	Kübler et al. (2020)
Product/service quality management	Predict user ratings from OCRs to monitor satisfaction	$n = 120{,}000$ electronics OCRs	Sentiment analysis of comments using recurrent neural networks (LSTM, GRU, Bi-LSTM, BiGRU)	Sachin et al. (2020)

Table 3 (Continued)

Decision area	Pursued goal	Collected UGC data	Applied method(s)	Reference
Product/service quality managemt	Identify important comments/attributes discussed	$n = 10{,}000$ camera and video game OCRs with helpfulness votes	Training a BERT model to predict OCR helpfulness, allowing firms and users to focus on important OCRs	Xu et al. (2020)
Brand management	Identify changes in relevant topics during brand crises	$n = 261{,}988$ Volkswagen/Burger King/Under Armour blog posts	Multiple latent changepoint topic modeling based on LDA applied to posts	Zhong and Schweidel (2020)
New product/service develoopmt	Identify improvement ideas for new products	$n = 10{,}000$ electronics, beauty, home, kitchen OCRs (text, rating)	Training a RNN to predict innovativeness based on comments coded by experts; application to all comments	Zhang et al. (2021a)
Product/service quality managemt	Identify important topics for customer satisfaction	16 datasets with product category OCRs w. comments, ratings	Word frequency analysis to distinguish high- and low-rated OCRs for each product category	Kim et al. (2022)
Product/service quality managemt	Identify pain points (customer concerns being addressable)	$n = 4{,}209{,}101$ tweets on 20 brands	Training a model to predict pain points based on comments coded by experts; application to all comments	Salminen et al. (2022)
Product/service quality managemt	Identify important comments/attributes discussed	$n = 10{,}000$ Yelp shopping OCRs with helpfulness votes	Training a BERT model to predict OCR helpfulness, allowing firms and users to focus on important OCRs	Bilal and Almazroi (2023)
New product/service develoopmt	Identify the most wanted features for a new air pump	$n = 20$ air pump OCRs with comments and ratings	Summarization of OCRs and idea generation using an LLM based chatbot (ChatGPT)	Bouschery et al. (2023)
Customer relationship managemt	Automate the customer complaint answer process	$n = 30$ TripAdvisor hospitality complaints by customers	Generation of responses to hospitality complaints according to prompted instructions by managers (ChatGPT)	Koc et al. (2023)
Product/service quality managemt	Identify relevant aspects for hotel guest satisfaction	$n = 8539$ TripAdvisor hotel OCRs	Extraction of aspects that impact hotel satisfaction using ChatGPT and topic modeling; importance calculation	Jeong and Jihwan (2024)
Product/service quality managemt	Improve usage of Metaverse platform services	$n = 17{,}136$ Metaverse OCRs	Topic modeling of OCRs to derive topics/attributes followed by regression analysis (usage intention)	Kumari et al. (2024)

Table 3 (Continued)

Decision area	Pursued goal	Collected UGC data	Applied method(s)	Reference
Product/service quality managemt	Identify relevant topics for hotel guest satisfaction	$n = 1{,}031{,}478$ TripAdvisor hotel OCRs	Topic modeling of OCRs using GPT topic and Falcon topic	Praveen et al. (2024)
Product/service quality managemt	Improve customer satisfaction with hotel services	$n = 96{,}322$ hotel OCRs (comment, rating)	Topic modeling of OCRs to derive topics/service attributes followed by regression analysis (satisfaction)	Zhang and Xu (2024)
Customer relationship managemt	Automate managerial responses to travel OCRs	$n = 446{,}663$ CTrain tour OCRs, 21% received a managerial response	Training a BERT model to categorize the response strategy to an OCR and neg. bin. regress. to predict adequacy	Zhang et al. (2024)
Product/service quality managemt	Improve platform and app acceptance (IKEA, others)	8 datasets with $n = 160{,}910$ platform and app OCRs	Training and applying a BERT model to predict acceptance based on comments coded by experts and ChatGPT	Baier et al. (2025)

For example, extending the already discussed regression analysis by Decker and Trusov (2010), Qi et al. (2016) as well as Xiao et al. (2016) applied a combination of conjoint analysis and the Kano model to OCRs to estimate the importance of product attributes from a customer's point of view. Rese et al. (2014) developed a lexicon-based approach to predict technology acceptance model (TAM) construct scores from OCR comments. This approach allows to monitor technology acceptance over time since OCRs have a time-stamp (the time at which they were written), an important asset when a new product, service, or app is introduced into a market and receives updates and relaunches over time. Kübler et al. (2020) also developed a time-dependent prediction model for brand awareness, impression, purchase intention, and satisfaction that can be used for (continuous) brand management. Timoshenko and Hauser (2019) trained a CNN to predict the informativeness of an OCR which helps to identify important OCRs on which to concentrate. Büschken and Allenby (2016) found in their LDA analysis of hotel OCRs topics like check-in problems, near-by attractions, or room cleanliness problems as well as corresponding customer segments that discussed one or more of these topics.

Recently, based on the Transformed-based deep learning methods discussed in the previous section, enormous methodological progress has been made. As discussed, today, OCRs can be easily analyzed using LLMs, such as BERT or ChatGPT. For example, Koc et al. (2023) applied ChatGPT to customer complaint management based on TripAdvisor posts. ChatGPT received instructions how to answer to these complaints. Then, a sample of 30 complaints and responses by ChatGPT and by trained hotel managers was shown to 40 industry experts who were asked to evaluate these answers. Concerning dimensions like credibility, apology, attentiveness, timeliness, redress, facilitation, and overall satisfactory, the evaluations of the chatbot's responses were significantly better rated.

Praveen et al. (2024) also analyzed TripAdvisor hotel OCRs, but applied LLM based topic models (using GPT and Falcon as LLMs) as well as ChatGPT and compared their results with analyses using traditional data analysis and machine learning. They found that discriminative and generative deep learning based on the Transformer architecture clearly outperformed the traditional methods and that the derived results for improving the hotel services were very helpful from a management perspective. Baier et al. (2025) extended the analysis by Rese et al. (2014) insofar that they replaced the lexicon-based approach to predict TAM construct scores used there by discriminative deep learning (multilingual BERT). They trained these models with product, service, and app OCRs rated by human experts and ChatGPT. The trained model predicts perceived informativeness, perceived enjoyment, perceived usefulness, perceived ease of use, attitude towards using, and behavioral intention to use over time in an adequate manner as could be demonstrated by comparisons with traditional TAM surveys. However, concerning the important decision areas price management and sales force management, we could not find—up to now—convincing UGC data collection and analysis use cases.

5.2 Image- and Video-Based Use Cases

The widespread adoption of social media platforms, such as Facebook, Instagram, and TikTok, as well as video-sharing platforms, such as YouTube, has significantly expanded the opportunities for users to curate and share content online, moving beyond written text to include images and videos (Klostermann et al. 2018). Moreover, established platforms for OCRs have long facilitated sharing pictures and videos, providing consumers with a platform to express their product and brand experiences in a more visually engaging way (Giglio et al. 2020; Kübler et al. 2024). This, in turn, generates valuable data for marketing researchers. Advances in data collection techniques and data processing methods for these multimedia formats have opened new possibilities for researchers. This has been particularly evident in the last few years, enabling them to explore consumer behavior analysis based on user-generated image, audio, and video content. Table 4 provides an overview of various use cases in marketing research articles that have analyzed user-generated images (photos, drawings) and videos to answer questions related to brand management, product/service quality management, and customer/market selection, among others.

In an early example, Baier et al. (2012) demonstrated that uploaded holiday, fashion, and apartment photos (during a survey or on social networks) can be used to group individuals into lifestyle segments based on the color distributions of the photos and other features. Similarly, Deng and Liu (2021) focused on segmentation, using over 14,000 Beijing-visit-related photos from Instagram. They employed facial and background recognition to extract features from uploaded images, derived, and discussed tourist segments with different preferences.

Instagram has become a particularly intriguing data source for both academic researchers and marketing managers, especially in the field of brand management. Studies by Liu et al. (2020) and Nanne et al. (2020) utilized photos uploaded to social media platforms that include references to brands in their analysis. Nanne et al. (2020) employed Google Cloud Vision API and other network approaches to categorize content, while Liu et al. (2020) used a pre-trained CNN to predict four attribute scores (glamorous, rugged, healthy, and fun) for each brand. Marketing managers can benefit from images posted in OCRs to better understand consumer experiences. This insight was highlighted by Zhang and Luo (2023) as well as Kübler et al. (2024), who utilized OCR images from Yelp and Amazon. Zhang and Luo (2023) focused on restaurant OCRs to assess the predictive power of uploaded photos in relation to a restaurant's success, while Kübler et al. (2024) emphasized the significance of images in OCRs in enhancing their helpfulness.

User-generated image content has much greater currency as a topic in marketing research than user-generated audio and video content. However, studies by Park et al. (2023) and Agrawal and Mittal (2024) emphasize the potential of video data in understanding consumer preferences. Agrawal and Mittal (2024) used product review videos from influencers on YouTube to extract features, and they conducted binomial regression analysis on likes and comments. Similarly, Park et al. (2023) analyzed product review videos on YouTube, extracting audio and video features for analysis. Their findings highlighted the influence of facial expressions, emotions, and voice pitch on the perceived helpfulness of the reviews.

Table 4 Overview of image- and video-based use cases of decision support in marketing management

Decision area	Pursued goal	UGC data source	Applied method(s)	Reference
Customer/market selection	Identify lifestyle segments	Posted (= preferred) holiday photos from $n = 478$ respondents (survey)	Clustering of images based on extracted features (e.g., color/texture/edge distribution, occurrence of objects)	Baier et al. (2012)
Brand management	Identify/monitor brand perception	$n = 10{,}375$ photos dealing with McDonald's posted on Instagram	Clustering of images based on extracted image labels using Google Cloud Vision API and caption texts	Klostermann et al. (2018)
Product/service quality managemt	Identify important hotel attributes in the luxury segment	$n = 7395$ photos of the interior of luxury hotels posted on TripAdvisor	Frequency analysis of extracted hotel attributes using Wolfram Mathematica software for image classification	Giglio et al. (2020)
Brand management	Identify influenceable reasons for user engagement	$n = 18{,}790$ tweets containing $n = 4537$ airline-related photos	Prediction of engagement based on extracted photo features by Google Cloud Vision and neg. bin. regression	Li and Xie (2020)
Brand management	Identify/monitor brand perception	$n = 114{,}367$ photos posted on Instagram dealing with 56 brands	Prediction of attribute scores (e.g., glamorous, rugged) based on a pre-trained CNN	Liu et al. (2020)
Brand management	Identify/monitor brand perception	$n = 21{,}738$ photos posted on Instagram dealing with 24 brands	Extraction of content categories using Google Cloud Vision API and other network approaches	Nanne et al. (2020)
Customer/market selection	Identify tourist segments	$n = 14{,}886$ Beijing-visit-related photos posted on Instagram	Extraction of contained faces/people, their age, gender, and other content categories, age-group comparisons	Deng and Liu (2021)
Brand management	Identify/monitor brand perception	$n = 4743$ brand-related collages posted on a visual elicitation platform	Extraction of content categories in the collages and clustering them brand related using guided LDA	Dzyabura and Peres (2021)
Brand management	Identify/monitor brand perception & purchase intention	$n = 214{,}563$ commented selfies with brand logo posted on Twitter	Extraction of brand names and selfie types using CNNs, prediction of purchase intention	Hartmann et al. (2021)
Brand management	Monitor competitor strategy	$n = 29{,}145$ photos containing vehicles, apparel, or faces (real or fake)	Extraction of object categories contained using deep learning included in cloud platforms (AWS, GCP, Azure)	Wang et al. (2022)
Brand management	Identify important visual features of OCRs	$n = 13{,}840$ product review videos on YouTube, helpfulness ratings	Extraction of audio and video features of the reviews, regression analysis to predict the stated review helpfulness	Park et al. (2023)

Table 4 (Continued)

Decision area	Pursued goal	UGC data source	Applied method(s)	Reference
Customer/market selection	Better understand motifs of sustainable fashion buyers	$n = 650$ Instagram posts using the hashtag #sustainablefashion	Frequency analysis of manually extracted image and text features (e.g., type of clothing, presence of persons)	Skinner et al. (2023)
Brand management	Monitor satisfaction with companies OCR-based	$n = 1{,}098{,}222$ images in Yelp OCRs coded as pos./neg. (star rating)	Finetuning ViT to predict sentiment scores (pos./neg.) based solely on the contained images in firm OCRs	Truong and Lauw (2023)
Brand management	Identify/monitor brand perception & purchase intention	$n = 755{,}758$ restaurant-related photos posted on Yelp, survival data	Extraction of photo content categories, prediction of restaurant survival based on the number and content of photos	Zhang and Luo (2023)
Brand management	Identify important product review video attributes	$n = 172$ product review videos on YouTube, posted by influencers	Extraction of video features through inspection and negative binomial regression of their likes and comments	Agrawal and Mittal (2024)
Brand management	Identify important product review attributes	$n = 97{,}947$ Amazon OCRs, including $n = 6060$ images	Tobit binomial regression to predict review helpfulness from the number of images, the review length, and ratings	Kübler et al. (2024)
Product/service quality managemt	Improve customer satisfaction with restaurant services	$n = 151{,}176$ restaurant-related Tripadvisor photos, star ratings for attributes	Extraction of photo content categories (quality indicators), prediction of star ratings based on these categories	Sharma et al. (2024)

A closer look at Table 4 shows that image and video analysis is mainly used in the decision area of brand management. Companies try to understand the context in which their brands are posted and whether the derived context features can be used for brand positioning. However, it should be noted that the applied methodologies might still be in a testing stage: Most of the use cases in Table 4 are probably further away from practice than the use cases in Table 3. Maybe, as discussed in Subsubsect. 4.2.4, the upcoming of improved LLMs and chatbots (e.g., Google's Transformer-based Gemini family, see Gemini Team 2024), that are able to summarize and analyze multimodal data sources in a convenient and valid manner, is here a huge step forward.

6 A Guideline for UGC Data Collection and Analysis for Researchers and Practitioners

The collection and analysis of UGC and other unstructured data is resource-consuming and has presented challenges for marketing researchers and practitioners (de Haan et al. 2024; Wedel and Kannan 2016). With the increasing availability of texts, images, videos, and other non-numeric forms of data, marketers can access a wealth of information previously unavailable. However, extracting actionable insights from these data types remains complex due to the lack of structured guidelines. Therefore, selecting adequate data sources and methods to solve a specific problem is essential.

To balance the potential benefits and the associated challenges of decision support based on unstructured data (problem resolution, costs of data collection and analysis, reliability and validity problems), de Haan et al. (2024) proposed a three-step approach as a guideline: (1) problem identification, (2) solution development, and (3) problem resolution. This process is grounded in organizational learning theory and aims to help managers choose the appropriate data sources and analysis methods based on their specific business needs. Therefore, as described below, we found that this three-step approach can be easily adapted to our context of UGC-based decision support in marketing management.

The first step, problem identification, requires the marketing managers to describe and characterize the problem for which decision support is needed. Here, the discussed decision areas and pursued UGC analysis goals from Subsects. 5.1 and 5.2 (see Tables 3 and 4) are helpful starting points. For many decision areas and goals, successful use cases of UGC data collection and analysis are discussed for customer/market selection, product/service quality management, or brand management. De Haan et al. (2024) argue in their guideline that in this first step, management should distinguish whether the primary analysis goal is an explorative one (e.g., brand management with a medium- to long-term business impact) or an exploitative one (e.g., new product/service management with a short- to medium-term business impact) since this characterization later affects the selection of an adequate data source.

The second step, solutions development, calls for marketing management to check which potential sources of UGC data are available for the identified problem: internal or external ones from the company's point of view, or even both.

- Internal data sources, such as OCRs and complaints stored on the company's website or in the company's customer database, can be accessed more easily. Especially for exploitative goals, like product/service quality management, such internal UGC data sources seem sufficient since they contain detailed descriptions of complaints and lead to improvement hints.
- External data sources, such as review sites or social networks (see Sect. 2), offer advantages for explorative goals since they contain in their OCR comments and posts attitudes and preferences concerning whole markets, not only the company's own brands, products, and services. This allows, e.g., insights on important product/service features from a general customer's point of view. However, collecting UGC data from these external sources is much more demanding since specific web scrapers or APIs for scraping text, images, and videos must be applied (see Sect. 3). Of course, in both cases, for internal and external data sources, specific data forms (text, image, audio, video, or a combination thereof) make data preprocessing necessary (see Subsects. 4.1 and 4.2).

The third step, problem resolution, is the final step in which marketing management decides which data to use and which methods to apply. They evaluate the potential solutions, weigh the costs and benefits, and implement the chosen strategy. De Haan et al. (2024) refer to the following questions that help selecting the best combination of data sources and methods (adapted to our UGC-based decision support):

- Which UGC data sources are available (internal data sources) and/or can be scraped (external sources)?
- How many UGC single data units are necessary to obtain robust insights?
- Which forms (text, image, audio, video, or a combination thereof) are available and which will be most appropriate?
- What are the privacy and legal boundaries for collecting external data and what are possible technical access requirements?
- What are the limits of data completeness and censorship?
- How timely should the collected UGC data units be?
- Which feature extraction method(s) is/are necessary, and can transfer models be applied?
- Is data aggregation necessary/useful, and if so, at what level?

Given the dynamic nature of unstructured data such as UGC data, marketing management must conduct ongoing assessments. New methods or data may emerge during the resolution phase, thus requiring continuous adaptation.

The three-step approach by de Haan et al. (2024) provides a structured way for marketing managers to utilize the potential of UGC and other unstructured data. When applied in combination with the insights from the presented paper, this strategy has the potential to become a powerful instrument for businesses, at least in consumer goods markets. It not only offers a structured roadmap for their actions but also furnishes them with a comprehensive array of data-gathering techniques, machine learning, and deep learning methodologies, as well as practical use cases exemplifying real-world applications of various data combinations and analyses.

Ultimately, this strategy will allow businesses to stay competitive by extracting meaningful insights from UGC.

7 Conclusions and Outlook

This paper presents a comprehensive review of the literature dealing with or contributing to UGC-based decision support. It covers data collection methods, their advantages and disadvantages, analysis approaches, including machine learning and deep learning methods, and various use cases that illustrate the wide range of questions, goals, and results in the marketing and management literature streams. The findings reveal that the structured integration of UGC into decision-making processes in marketing management can potentially offer significant benefits across various areas, such as customer and market selection, product and service quality management, as well as brand perception monitoring.

Rapid advancements in NLP technologies, particularly the use of LLMs such as BERT and ChatGPT, have democratized access to UGC analysis. These tools allow even non-experts to collect, analyze, and derive valuable insights from textual data, offering significant advantages in terms of efficiency and accessibility. However, while LLMs are well-suited for text-based UGC, image and video content continue to require specialized skills and expertise. Despite the valuable insights and practical models offered by academic research for marketing professionals, the dynamic nature of technology, especially UGC, requires forward-thinking solutions to address existing and future challenges. Furthermore, LLMs face technical challenges, such as the need for vast computational resources and the potential misinterpretation of complex data (e.g., images or ironic texts). Future review articles could focus on these limitations, evaluating the effectiveness and scalability of different models and proposing solutions to reduce computational costs and improve the interpretability of UGC data.

Looking to the future, we anticipate several developments in the field of UGC analysis. First, as natural language processing models continue to improve, we expect even more sophisticated sentiment analysis and topic modeling capabilities to emerge, which will enhance decision-making based on OCR data for marketing managers. Second, advancements in computer vision and machine learning methods for image and video data will likely reduce the reliance on human expertise, thus enabling more automated and scalable analysis. This can be crucial for marketing teams seeking to leverage UGC in real-time environments, such as social media platforms.

Another promising avenue for further research is the integration of UGC analysis with other data sources, such as structured sales data and customer feedback. By combining these data streams, marketing teams can gain a more holistic understanding of customer behavior and preferences, leading to more informed and effective decision-making. The ability to synthesize insights across multiple data types will be a critical capability for marketing teams in the coming years.

It is essential to underline that ethical considerations and privacy concerns will become increasingly important as the practice of collecting UGC data grows. Prac-

titioners and researchers must navigate data protection regulations such as GDPR, which impose strict requirements on the use of personal data, even when such content is publicly available. Future research could explore how to balance the need for insightful data analysis with the objective of protecting user privacy and ensuring ethical data collection practices. Future review articles could focus on the ongoing evolution of these regulations and their impact on UGC analysis, exploring, for example, how different regions handle consent and privacy and the best practices for researchers and marketers to stay compliant with such laws.

Finally, a key restriction mentioned is the challenge of ensuring the representativeness of UGC data collected through APIs and web scraping. Future studies could review the extent to which biases are introduced through non-representative sampling, such as selective access to certain types of UGC (e.g., data from specific platforms or demographic groups). Research could investigate methods to mitigate these biases and improve the validity of marketing insights drawn from UGC.

To conclude, while significant progress has been made in the field of UGC data collection and analysis, ongoing advancements in machine learning, natural language processing, and multimedia analysis hold the potential to further enhance the effectiveness of UGC-based decision-making in marketing. Continued innovation in these areas will ensure that UGC remains a valuable and versatile asset for marketers, enabling them to respond more effectively to changing customer needs and market conditions.

Conflict of interest D. Baier, R. Decker and Y. Asenova declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ackley, David H., Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for Boltzmann machines. *Cognitive Science* 9(1):147–169. https://doi.org/10.1016/S0364-0213(85)80012-4.

Agrawal, Shiv R., and Divya Mittal. 2024. Optimizing marketing strategy: a video analysis approach. *Marketing Intelligence & Planning* https://doi.org/10.1108/MIP-12-2023-0655.

Amuhda, S. 2017. Web crawler for mining web data. *International Research Journal of Engineering and Technology* 4(2):128–136.

Bahdanau, Dzmitry, Cho Kyunghyun, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate, arXiv:1409.0473.http://arxiv.org/pdf/1409.0473v7.

Baier, Daniel, Ines Daniel, Sarah Frost, and Robert Naundorf. 2012. Image data analysis and classification in marketing. *Advances in Data Analysis and Classification* 6(4):253–276. https://doi.org/10.1007/s11634-012-0116-0.

Baier, Daniel, Andreas Karasenko, and Alexandra Rese. 2025. Measuring technology acceptance over time through transfer models based on online customer reviews. *Journal of Retailing and Consumer Services*, accepted.

Balasubramanian, Sridhar, and Vijay Mahajan. 2001. The economic leverage of the virtual community. *International Journal of Electronic Commerce* 5(3):103–138. https://doi.org/10.1080/10864415.2001.11044212.

Balducci, Bitty, and Detelina Marinova. 2018. Unstructured data in marketing. *Journal of the Academy of Marketing Science* 46(4):557–590. https://doi.org/10.1007/s11747-018-0581-x.

Bayindir, Nisa, and Erik Winther Paisley. 2019. Brand discovery: Examining the ways digital consumers discover new brands, products, and services: GWI Insights Report. https://www.gwi.com/hubfs/Downloads/Brand_Discovery-2019.pdf. Accessed 24 July 2024.

Bergman, Jesper, and Oliver B. Popov. 2023. Exploring dark web crawlers: A systematic literature review of dark web crawlers and their implementation. *IEEE Access* 11:35914–35933. https://doi.org/10.1109/access.2023.3255165.

Bilal, Muhammad, and Abdulwahab Ali Almazroi. 2023. Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews. *Electronic Commerce Research* 23(4):2737–2757.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of machine Learning research* 3(1):993–1022.

Boegershausen, Johannes, Hannes Datta, Abhishek Borah, and Andrew T. Stephen. 2022. Fields of gold: Scraping web data for marketing insights. *Journal of Marketing* 86(5):1–20. https://doi.org/10.1177/00222429221100750.

Bouschery, Sebastian G., Vera Blazevic, and Frank T. Piller. 2023. Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management* 40(2):139–153. https://doi.org/10.1111/jpim.12656.

Brooke, Connor. 2024. Trustpilot vs PowerReviews: Which is more effective with customer reviews? https://www.business2community.com/consumer-marketing/trustpilot-vs-powerreviews-which-is-more-effective-with-customer-reviews-02087412. Accessed 24 July 2024.

Büschken, Joachim, and Greg M. Allenby. 2016. Sentence-based text analysis for customer reviews. *Marketing Science* 35(6):953–975. https://doi.org/10.1287/mksc.2016.0993.

Chang, Yupeng, Xu Wang, Wang Jindong, Yuan Wu, Yang Linyi, Zhu Kaijie, Hao Chen, Yi Xiaoyuan, Wang Cunxiang, Wang Yidong, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* https://doi.org/10.1145/3641289.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv:1406.1078.

Chollet, François. 2021. *Deep learning with Python*, 2nd edn., New York: Manning.

Daniel, Ines. 2014. *Validität von Lebensstilsegmentierungen aufgrund der Bewertung vorgegebener digitaler Bilder*. Wiesbaden: Springer Gabler.

Darmawan, Irfan, Muhamad Maulana, Rohmat Gunawan, and Nur Widiyasono. 2022. Evaluating web scraping performance using XPath, CSS selector, regular expression, and HTML DOM with multiprocessing technical applications. *JOIV International Journal on Informatics Visualization* 6(4):904–910. https://doi.org/10.30630/joiv.6.4.1525.

de Haan, Evert, Manjunath Padigar, Siham El Kihal, Raoul Kübler, and Jaap E. Wieringa. 2024. Unstructured data research in business: Toward a structured approach. *Journal of Business Research* 177:114655. https://doi.org/10.1016/j.jbusres.2024.114655.

Decker, Reinhold, and Michael Trusov. 2010. Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing* 27(4):293–307. https://doi.org/10.1016/j.ijresmar.2010.09.001.

Dellarocas, Chrysanthos, Zsolt Katona, and William Rand. 2013. Media, aggregators, and the link economy: Strategic hyperlink formation in content networks. *Management Science* 59(10):2360–2379. https://doi.org/10.1287/mnsc.2013.1710.

Deng, Ning, and Jiayi Liu. 2021. Where did you take those photos? Tourists' preference clustering based on facial and background recognition. *Journal of Destination Marketing & Management* 21:100632. https://doi.org/10.1016/j.jdmm.2021.100632.

Deshmukh, Smita, and Kantilal Vishwakarma. 2021. A survey on crawlers used in developing search engine. In *Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1446–1452. IEEE. May 6–8, 2021.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805.

Dhenakaran, S.S., and K. Thirugnana Sambanthan. 2011. Web crawler: An overview. *International Journal of Computer Science and Communication* 2(1):265–267.

Dorner, Verena, Marcus Giamattei, and Matthias Greiff. 2020. The market for reviews: Strategic behavior of online product reviewers with monetary incentives. *Schmalenbach Business Review* 72(3):397–435. https://doi.org/10.1007/s41464-020-00094-y.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint: 2010.11929.

Duarte, Vannessa, Sergio Zuniga-Jara, and Sergio Contreras. 2022. Machine learning and marketing: A systematic literature review. *IEEE Access* 10:93273–93288. https://doi.org/10.1109/access.2022.3202896.

Dwivedi, Y.K., N. Kshetri, L. Hughes, E.L. Slade, A. Jeyaraj, A.K. Kar, A.M. Baabdullah, A. Koo-hang, V. Raghavan, M. Ahuja, H. Albanna, M.A. Albashrawi, A.S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, L. Brooks, D. Buhalis, L. Carter, and R. Wright. 2023. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71:102642.

Dzyabura, Daria, and Renana Peres. 2021. Visual elicitation of brand perception. *Journal of Marketing* 85(4):44–66. https://doi.org/10.1177/0022242921996661.

Egger, Roman, Markus Kroner, and Andreas Stöckl. 2022. Web scraping collecting and retrieving data from the web. In *Applied data science in tourism—interdisciplinary approaches, methodologies, and applications*, ed. Roman Egger, 67–82. Heidelberg: Springer.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39(11):27–34. https://doi.org/10.1145/240455.240464.

Feldman, Ronen, and Igo Dagan. 1995. Knowledge discovery in textual databases. In *Proceedings of KDD-95: The First International Conference on Knowledge Discovery & Data Mining, Montreal, August 20–21, 1995*, ed. Usama Fayyad, 112–117. Menlo Park: AAAI.

Gemini Team. 2024. *Gemini: A family of highly capable multimodal models*. arXiv preprint arXiv:2312.11805.

George, S., and V. Srividhya. 2022. Automated summarization of restaurant reviews using hybrid approaches. *ICTACT Journal on Soft Computing* 12(4):2690–2696.

Gezici, Bahar, Necva Bölücü, Ayca Tarhan, and Burcu Can. 2019. Neural sentiment analysis of user reviews to predict user ratings. In *2019 4th International conference on computer science and engineering (UBMK)*, 629–634. IEEE.

Giglio, Simona, Eleonora Pantano, Eleonora Bilotta, and T.C. Melewar. 2020. Branding luxury hotels: Evidence from the analysis of consumers' "big" visual data on TripAdvisor. *Journal of Business Research* 119:495–501. https://doi.org/10.1016/j.jbusres.2019.10.053.

Gupta, Anish, and Priya Anand. 2015. Focused web crawlers and its approaches. In *1st International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, 619–622. IEEE.

Guyt, Jonne Y., Hannes Datta, and Johannes Boegershausen. 2024. Unlocking the potential of web data for retailing research. *Journal of Retailing* 100(1):130–147. https://doi.org/10.1016/j.jretai.2024.02.002.

Hartmann, Jochen, Mark Heitmann, Christina Schamp, and Oded Netzer. 2021. The power of brand selfies. *Journal of Marketing Research* 58(6):1159–1177. https://doi.org/10.1177/00222437211037258.

Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing* 40(1):75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005.

Herhausen, Dennis, Stefan F. Bernritter, Eric W. Ngai, Ajay Kumar, and Dursun Delen. 2024. Machine learning in marketing: Recent progress and future research directions. *Journal of Business Research* 170:114254. https://doi.org/10.1016/j.jbusres.2023.114254.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Humphreys, Ashley, and Rebecca Jen-Hui Wang. 2018. Automated text analysis for consumer research. *Journal of Consumer Research* 44(6):1274–1306.

Hutto, C.J., and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media. University of Michigan, Ann Arbor, June 1–4, 2014*.

Islam, Rakibul, and Minhaz F. Zibran. 2018. A comparison of software engineering domain specific sentiment analysis tools. In *25th IEEE International Conference on Software Analysis, Evolution and Reengineering, Campobasso, Italy*, 487–491.

Jeong, Nayoung, and Lee Jihwan. 2024. An aspect-based review analysis using ChatGPT for the exploration of hotel service failures. *Sustainability* 16(4):1640.

Jovanovic, Mladan, and Mark Campbell. 2022. Generative artificial intelligence: Trends and prospects. *IEEE Computer (Computer)* 55(10):107–112. https://doi.org/10.1109/mc.2022.3192720.

Khder, Moaiad. 2021. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and its Applications* 13(3):145–168. https://doi.org/10.15849/ijasca.211128.11.

Kim, Siung, and Nammee Moon. 2024. Vision Transformer based classification of bulky waste. In *Advances in Computer Science and Ubiquitous Computing—International Conference on Computer Science and its Applications and the International Conference on Ubiquitous Information Technologies and Applications*, ed. Ji Su Park, Laurence T. Yang, Yi Pan, and James L. Parks, 438–442. Singapore: Springer Nature.

Kim, Taeyong, Hwang Seungsoo, and Minkyung Kim. 2022. Text analysis of online customer reviews for products in the FCB quadrants: Procedure, outcomes, and implications. *Journal of Business Research* 150:676–689. https://doi.org/10.1016/j.jbusres.2022.05.077.

Klostermann, Jan, Anja Plumeyer, Daniel Böger, and Reinhold Decker. 2018. Extracting brand information from social networks: Integrating image, text, and social tagging data. *International Journal of Research in Marketing* 35(4):538–556. https://doi.org/10.1016/j.ijresmar.2018.08.002.

Koc, Erdogan, Sercan Hatipoglu, Oguzhan Kivrak, Cemal Celik, and Kaan Koc. 2023. Houston, we have a problem!: The use of ChatGPT in responding to customer complaints. *Technology in Society* 74:102333.

Kramer, Mark A. 1991. Nonlinear principal component analysis Using autoassociative neural networks. *American Institute of Chemical Engineers Journal* 37(2):233–243. https://doi.org/10.1002/aic.690370209.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60(6):84–90. https://doi.org/10.1145/3065386.

Kübler, Raoul V., Anatoli Colicev, and Koen H. Pauwels. 2020. Social media's impact on the consumer mindset: When to use which sentiment extraction tool? *Journal of Interactive Marketing* 50(1):136–155. https://doi.org/10.1016/j.intmar.2019.08.001.

Kübler, Raoul V., Lara Lobschat, Lina Welke, and Hugo van der Meij. 2024. The effect of review images on review helpfulness: A contingency approach. *Journal of Retailing* 100(1):5–23. https://doi.org/10.1016/j.jretai.2023.09.001.

Kumari, Vandana, K. Bala Pradip, and Shibashish Chakraborty. 2024. A text mining approach to explore factors influencing consumer intention to use Metaverse platform services: Insights from online customer reviews. *Journal of Retailing and Consumer Services* 81:103967. https://doi.org/10.1016/j.jretconser.2024.103967.

Levene, Mark, and Alexandra Poulovassilis. 2001. Web dynamics. *Software Focus* 2(2):60–67. https://doi.org/10.1002/swf.30.

Li, Yiyi, and Ying Xie. 2020. Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research* 57(1):1–19. https://doi.org/10.1177/0022243719881113.

Li, Shugang, Fang Liu, Zhang Yuqi, Zhu Boyi, He Zhu, and Zhaoxu Yu. 2022. Text mining of user-generated content (UGC) for business applications in e-commerce: A systematic review. *Mathematics* 10(19):3554. https://doi.org/10.3390/math10193554.

Liu, Liu, Daria Dzyabura, and Natalie Mizik. 2020. Visual listening in: Extracting brand image portrayed on social media. *Marketing Science* 39(4):669–686. https://doi.org/10.1287/mksc.2020.1226.

Liu, Yuanshao, Wang Junqi, and Xiaolong Wang. 2018. Learning to recognize opinion targets using recurrent neural networks. *Pattern Recognition Letters* 106:41–46.

Lomborg, Stine, and Anja Bechmann. 2014. Using APIs for data collection on social media. *The Information Society—An International Journal* 30(4):256–265. https://doi.org/10.1080/01972243.2014.915276.

Ma, Liye, and Baohong Sun. 2020. Machine Learning and AI in marketing—Connecting computing power to human insights. *International Journal of Research in Marketing* 37(3):481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005.

Maleshkova, Maria, Carlos Pedrinaci, and John Domingue. 2010. Investigating web APIs on the world wide web. In *8th IEEE European Conference on Web Services, December 1–3, 2010*, 107–114. Ayia Napa: IEEE.

Manias, George, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, and Dimosthenis Kyriazis. 2023. Multilingual text categorization and sentiment analysis: A comparative analysis of the utilization of multilingual approaches for classifying Twitter data. *Neural computing & applications* 35(29):21415–21431. https://doi.org/10.1007/s00521-023-08629-3.

Menczer, Filippo, Gautam Pant, and Padmini Srinivasan. 2004. Topical web crawlers; Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)* 4(4):378–419. https://doi.org/10.1145/1031114.1031117.

Naab, Teresa N., and Annika Sehl. 2017. Studies of user-generated content: A systematic review. *Journalism* 18(10):1256–1273.

Nakayama, Atsuho, and Daniel Baier. 2020. Predicting brand confusion in imagery markets based on deep learning of visual advertisement content. *Advances in Data Analysis and Classification* 14(4):927–945. https://doi.org/10.1007/s11634-020-00429-0.

Nanne, Annemarie J., Marjolijn L. Antheunis, Chris G. van der Lee, Eric O. Postma, Sander Wubben, and Guda van Noort. 2020. The use of computer vision to analyze brand-related user generated image content. *Journal of Interactive Marketing* 50(1):156–167. https://doi.org/10.1016/j.intmar.2019.09.003.

Ngai, Eric W., and Yuanyuan Wu. 2022. Machine Learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research* 145:35–48. https://doi.org/10.1016/j.jbusres.2022.02.049.

Open, A. 2023. GPT-4 technical report. arXiv. https://arxiv.org/pdf/2303.08774.

Open, A. 2024. Introducing ChatGPT. https://openai.com/index/chatgpt/. Accessed 8 Jan 2025.

Paget, Sammy. 2024. Local consumer review survey 2024: Trends, behaviors, and platforms explored. https://www.brightlocal.com/research/local-consumer-review-survey/. Accessed 21 July 2024.

Park, Kyungmin, Stephanie Lee, Shahryar Doosti, and Yong Tan. 2023. Provision of helpful review videos: Effects of video characteristics on perceived helpfulness. *Production and Operations Management* 32(7):2031–2048. https://doi.org/10.1111/poms.13969.

Parmar, Niki, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, July 10–15, 2018*, ed. Jennifer Dy, Andreas Krause, 4055–4064.

Pennebaker, James W., E. Francis Martha, and Roger J. Booth. 2001. *Linguistic inquiry and word count*. LIWC, Vol. 2001. Mahwah: Erlbaum.

Praveen, S.V., Pranshav Gajjar, Rajeev Kumar Ray, and Ashutosh Dutt. 2024. Crafting clarity: Leveraging large language models to decode consumer reviews. *Journal of Retailing and Consumer Services* 81:103975.

Puschmann, Cornelius, and Julian Ausserhofer. 2017. Social data APIs: origin, types, issues. In *The datafied society*, ed. M.T. Schäfer, K. Es, 147–154. Amsterdam University Press.

Qi, Jiayin, Zhang Zhenping, Jeon Seongmin, and Yanquan Zhou. 2016. Mining customer requirements from online reviews: A product improvement perspective. *Information & Management* 53(8):951–963. https://doi.org/10.1016/j.im.2016.06.002.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Zhou Yanqi, Wei Li, and Peter J. Liu. 2019. *Exploring the limits of transfer learning with a unified text-to-text transformer*. arxiv:1910.10683.

Rese, Alexandra, Stefanie Schreiber, and Daniel Baier. 2014. Technology Acceptance Modeling of augmented reality at the point of sale: Can surveys be replaced by an analysis of online reviews? *Journal of Retailing and Consumer Services* 21(5):869–876. https://doi.org/10.1016/j.jretconser.2014.02.011.

Roberts, John H., Ujwal Kayande, and Stefan Stremersch. 2014. From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing* 31(2):127–140. https://doi.org/10.1016/j.ijresmar.2013.07.006.

Ruelens, Anna. 2022. Analyzing user-generated content using natural language processing: A case study of public satisfaction with healthcare systems. *Journal of Computational Social Science* 5(1):731–749. https://doi.org/10.1007/s42001-021-00148-2.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1987. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of*

cognition—Foundations, ed. David E. Rumelhardt, James L. McClelland, 318–362. Cambridge: MIT Press.

Sachin, Sharat, Abha Tripathi, Navya Mahajan, Shivani Aggarwal, and Preeti Nagrath. 2020. Sentiment analysis using gated recurrent neural networks. *SN Computer Science* 1:1–13.

Salminen, Joni, Mekhail Mustak, Juan Corporan, Jung Soon-gyo, and Bernard J. Jansen. 2022. Detecting pain points from user-generated social media posts using machine learning. *Journal of Interactive Marketing* 57(3):517–539. https://doi.org/10.1177/10949968221095556.

Sarker, Iqbal H. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science* 2(6):1–20.

Schermerhorn, Vaughn. 2023. How amazon continues to improve customer review experience with generative AI. https://www.aboutamazon.com/news/amazon-ai/amazon-improves-customer-reviews-with-generative-ai. Accessed 24 July 2024.

Sharkey, Amanda, Balázs Kovács, and Greta Hsu. 2023. Expert critics, rankings, and review aggregators: The changing nature of intermediation and the rise of markets with multiple intermediaries. *Academy of Management Annals* 17(1):1–36. https://doi.org/10.5465/annals.2021.0025.

Sharma, Ujjwal, Stevan Rudinac, Joris Demmers, Willemijn van Dolen, and Marcel Worring. 2024. From pixels to perceptions: Capturing high-level abstract concepts in visual user-generated content. *International Journal of Information Management Data Insights* 4(2):100269. https://doi.org/10.1016/j.jjimei.2024.100269.

Simonyan, Karen, and Andrew Zisserman. 2015. *Very deep convolutional networks for large-scale image recognition*. arXiv preprint: 1409.1556.

Skinner, Erin A., Brianna Le Busque, Jillian Dorrian, and Carla A. Litchfield. 2023. sustainablefashion on Instagram: A content and network analysis of user-generated posts. *Journal of Consumer Behaviour* 22(5):1096–1111. https://doi.org/10.1002/cb.2182.

Sutskever, Ilya, Oriol Vinyals, and V. Le Quoc. 2015. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: 28th Annual Conference on Neural Information Processing Systems 2014; December 8–13, 2014, Montreal, Canada*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, 3104–3112. Red Hook: Curran.

Tabassum, Ayisha, and Rajendra R. Patil. 2020. A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology* 7(6):4864–4867.

Tealab, Ahmed. 2018. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal* 3(2):334–340.

Thenmalar, S., and V.T. Geetha. 2011. Concept based focused crawling using ontology. *International Journal of Computer Applications* 26(7):29–32. https://doi.org/10.5120/3115-4282.

Timoshenko, Artem, and John R. Hauser. 2019. Identifying customer needs from user-generated content. *Marketing Science* 38(1):1–20. https://doi.org/10.1287/mksc.2018.1123.

Truong, Quoc-Tuan, and Hady W. Lauw. 2023. Concept-oriented transformers for visual sentiment analysis. In *Proceedings of the sixteenth ACM international conference on web search and data mining, WSDM 2023, February 27-March 3, 2023, Singapore*, 1111–1119.

Uppalapati, Padma J., Madhavi Dabbiru, and K.V. Rao. 2023. A comprehensive survey on summarization techniques. *SN Computer Science* 4(5):560. https://doi.org/10.1007/s42979-023-02007-5.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Jones Llion, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *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*, ed. Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, Rob Fergus, S.V.N. Vishwanathan, and Roman Garnett. Red Hook: Curran Associates.

Vickery, Graham, and Sacha Wunsch-Vincent. 2007. Participative web and user-created content: Web 2.0, wikis and social networking. https://www.oecd-ilibrary.org/science-and-technology/participative-web-and-user-created-content_9789264037472-en. Accessed 24 July 2024.

Wang, Jing, Min Weiqing, Hou Sujuan, Ma Shengnan, Zheng Yuanjie, and Shuqiang Jiang. 2022. LogoDet-3K: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18(1):1–19. https://doi.org/10.1145/3466780.

Wang, Zhengjue, Duan Zhibin, Hao Zhang, Wang Chaojie, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Stroudsburg, PA, USA, November, 2020*, ed. Bonnie Webber, Trevor Cohn, and He Yang Liu Yulan. Stroudsburg: Association for Computational Linguistics.

Wedel, Michel, and P.K. Kannan. 2016. Marketing analytics for data-rich environments. *Journal of Marketing* 80(6):97–121. https://doi.org/10.1509/jm.15.0413.

Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data* 3(1):1–40. https://doi.org/10.1186/s40537-016-0043-6.

Widyassari, Adhika P., Supriadi Rustad, Guruh F. Shidik, Edi Noersasongka, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences* 34(4):1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006.

Xiao, Geoffrey. 2020. Bad bots: Regulating the scraping of public personal information. *Harvard Journal of Law & Technology* 34(2):702–732.

Xiao, Shengsheng, Wei Chih-Ping, and Ming Dong. 2016. Crowd intelligence: Analyzing online product reviews for preference measurement. *Information & Management* 53(2):169–182. https://doi.org/10.1016/j.im.2015.09.010.

Xu, Hongyan, Liu Hongtao, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Pre-trained personalized review summarization with effective salience estimation. In *Findings of the Association for Computational Linguistics (ACL 2023), Toronto, Canada, July, 2023*, ed. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, 10743–10754. Stroudsburg: Association for Computational Linguistics.

Xu, Shuzhe, Salvador E. Barbosa, and Don Hong. 2020. BERT feature based model for predicting the helpfulness scores of online customers reviews. In *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*, 270–281. Cham: Springer.

Yadav, Vandana, Parul Verma, and Vinodini Katiyar. 2023. Long Short Term Memory (LSTM) model for sentiment analysis in social data for e-commerce products reviews in Hindi languages. *International Journal of Information Technology* 15(2):759–772.

Yang, Mochen, Ren Yuqing, and Gediminas Adomavicius. 2019. Understanding user-generated content and customer engagement on Facebook business pages. *Information Systems Research* 30(3):839–855. https://doi.org/10.1287/isre.2019.0834.

Zhang, Chenxi, and Zeshui Xu. 2024. Gaining insights for service improvement through unstructured text from online reviews. *Journal of Retailing and Consumer Services* 80:103898. https://doi.org/10.1016/j.jretconser.2024.103898.

Zhang, Mengxia, and Lan Luo. 2023. Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. *Management Science* 69(1):25–50. https://doi.org/10.1287/mnsc.2022.4359.

Zhang, Min, Brandon Fan, Ning Zhang, Wang Wenjun, and Weiguo Fan. 2021a. Mining product innovation ideas from online reviews. *Information Processing & Management* 58(1):102389. https://doi.org/10.1016/j.ipm.2020.102389.

Zhang, Xueying, Jiang Yunjiang, Yue Shang, Cheng Zhaomeng, Chi Zhang, Fan Xiaochuan, Yun Xiao, and Bo Long. 2021b. DSGPT: Domain-specific generative pre-training of transformers for text generation in e-commerce title and review summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, July 111–15, 2021*, ed. Fernando Diaz, Chirag Shah, 2146–2150. New York: ACM.

Zhang, Xin, Lei La, Guo Qiong I. Huang, and Xie Haoxiang. 2024. Strategies and conditions for crafting managerial responses to online reviews. *Tourism Management* 103:104911.

Zhao, Qingjuan, Niu Jianwei, and Xuefen Liu. 2022. ALS-MRS: Incorporating aspect-level sentiment for abstractive multi-review summarization. *Knowledge-Based Systems* 258:109942. https://doi.org/10.1016/j.knosys.2022.109942.

Zhong, Ning, and David A. Schweidel. 2020. Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science* 39(4):827–846. https://doi.org/10.1287/mksc.2019.1212.