mathematics

Review

Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review

Shugang Li [1], Fang Liu [1], Yuqi Zhang [1,*], Boyi Zhu [1], He Zhu [1] and Zhaoxu Yu [2]

[1] School of Management, Shanghai University, Shanghai 200444, China
[2] Department of Automation, East China University of Science and Technology, Shanghai 200237, China
* Correspondence: yuqi_zhang@shu.edu.cn

Abstract: In the Web2.0 era, user-generated content (UGC) provides a valuable source of data to aid in understanding consumers and driving intelligent business. Text mining techniques, such as semantic analysis and sentiment analysis, help to extract meaningful information embedded in UGC. However, research on text mining of UGC for e-commerce business applications involves interdisciplinary knowledge, and few studies have systematically summarized the research framework and application directions of related research in this field. First, based on e-commerce practice, in this study, we derive a general framework to summarize the mainstream research in this field. Second, widely used text mining techniques are introduced, including semantic and sentiment analysis. Furthermore, we analyze the development status of semantic analysis in terms of text representation and semantic understanding. Then, the definition, development, and technical classification of sentiment analysis techniques are introduced. Third, we discuss mainstream directions of text mining for business applications, ranging from high-quality UGC detection and consumer profiling, to product enhancement and marketing. Finally, research gaps with respect to these efforts are emphasized, and suggestions are provided for future work. We also provide prospective directions for future research.

Keywords: text mining; user-generated content (UGC); semantic analysis; sentiment analysis; business applications; consumer profiling

MSC: 91

Citation: Li, S.; Liu, F.; Zhang, Y.; Zhu, B.; Zhu, H.; Yu, Z. Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review. *Mathematics* **2022**, *10*, 3554. https://doi.org/10.3390/math10193554

Academic Editor: Radu Tudor Ionescu

Received: 9 September 2022
Accepted: 25 September 2022
Published: 29 September 2022

1. Introduction

In the digital era, the prevalence of the Internet and the development of information technology (IT) have empowered every user of the web to generate, share, and propagate content, which has transformed the contents of the web from marketer- to user-created content, namely user-generated content (UGC) [1]. Every user can easily access UGC generated by others via the Internet in real time. Nowadays, the Internet integrates various types of platforms, such as social network platforms (e.g., Sina Weibo, Twitter, and Instagram), cyber communication (e.g., Wechat, e-mail), online communities (e.g., Zhihu and Quora), electronic commerce (e.g., Amazon, Taobao, Meituan, and Little Red Book), and Internet finance (e.g., PayPal and Alipay). On electronic commerce (e-commerce) platforms, consumers actively generate an overwhelming amount of unstructured data within comparatively short timeframes with respect to interesting topics, reviews, and opinions. The considerable UGC, coupled with data about individual details, footprints, and behaviors, referred to as "big data", has provided extensive opportunities and challenges for researchers and practitioners.

By serving as a valuable and credible source of information for consumers throughout the purchasing lifecycle, UGC has transformed the ways in which consumer retrieve, visit, evaluate, select, and share experiences [2]. The impact of user-generated data on the consumer has been magnified. UGC is able to attract consumers' attention, drive consumer activities, reduce uncertainty and risk for consumers with respect to decision making, and encourage consumers to purchase [3]. These data, especially text content, embrace consumers' thoughts, opinions, and feelings regarding products and provide firms the opportunity to listen to consumers to get better understand them [4]. In addition, recent research argued that UGC is a fundamental prerequisite for devising new ideas and intuitions in terms of business intelligence. Despite being valuable for enterprises, massive and unstructured UGC created at an overwhelming rate, is of low value density. Consequently, an appropriate and effective method for extraction of knowledge embedded in text data is invaluable.

With the development of natural language processing (NLP) and deep learning techniques, UGC has gradually replaced consumer surveys as the critical method to understand consumers, consequently contributing to the improvement and innovation of products, services, and brands [5]. Extant research has also been conducted to investigate technological innovation on the basis of UGC analysis of topics, sentiments, opinions, etc.

UGC mining and applications are academic topics that involve interdisciplinary domains. The coupling of linguistics, consumer behavior, economics, psychology, and computer science is associated with challenges for scholars from different fields [6]. Given the interdisciplinary nature, few studies have systematically summarized the frameworks and application directions of related research and literature. Furthermore, multidisciplinary convergence increases the difficulty of presenting the theoretical background and technical foundations in-depth [7]. Therefore, in this study, we review recent research on UGC text mining techniques and their business applications in the field of e-commerce, providing a general framework and wide perspective of this area and emphasizing the challenges that need to be addressed in future research.

The remainder of this paper is organized as follows. In Section 2, we identify the literature collection scope and provide a general framework for related research. In Section 3, we present the sources and types of UGC in e-commerce practice. In Section 4, we introduce mainstream techniques for text mining. In Section 5, we summarize some common applications in the e-commerce field involving text mining of UGC. Finally, in Section 6, we discuss the highlights and challenges in extant research and suggest possible directions for future research.

2. Literature Identification and Collection

2.1. General Overview and Thematic Analysis

The aim of this research is to systematically review research on text mining based on UGC from the perspectives of various scientific domains. In order to adequately address this research problem, we consider the research fields of linguistics, computer science, and consumer behavior, with a focus on analysis of the relevant literature on UGC data in the e-commerce market. Accordingly, the background theme of this study is constructed to develop a robust comprehension of this research problem.

2.1.1. Literature Selection and General Overview

To select the most relevant publications, a search was executed using the following query in the Web of Science database and EI's Engineering village database:

('review*' OR 'comment*' OR '*wom' OR '*word?of?mouth' OR 'UGC' OR 'user-generated content*') AND ('consumer*' OR 'customer*' OR 'user*' OR 'traveler*' OR 'buyer*') AND ("text mining" OR "natural language process" OR 'NLP' or 'sentiment*' OR 'opinion' OR 'topic model*' OR 'semantic') AND ('e-commerce' OR "electronic commerce" OR 'social commerce' OR 'purchase' OR 'buying').

This query includes three main parts. (1) UGC is the objective of the research. In this query, keywords such as consumer and user were included to emphasize that business intelligence and knowledge should be consumer-oriented and based on the content created by individual consumers. (2) Methods related to text mining should be innovatively incorporated and even improved to understand textual UGC. (3) The research domain should lie in the e-commerce field and be restricted to product/service purchases. Conventionally, e-commerce refers to the activity of electronically purchasing or selling over the Internet. Social commerce, which allows consumers to engage in the sales and marketing of products and services [8], is often considered a subset of e-commerce [9]. Other Internet-based behaviors, such as video consumption behavior on video websites and message-forwarding behavior on social networking sites, fall outside the scope of this study.

We collected a total of 2413 and 2335 records from the Web of Science database and the EI database, respectively. After removing duplicate, invalid, and irrelevant records, a total of 3392 relevant articles were obtained. Invalid records refer to non-research or non-review paper records, including abstract records of conferences or lectures. Irrelevant records include but are not limited to the following: research papers that draw mainly on simulation modeling rather than empirical examination of real UGC datasets; research papers that do not investigate UGC text and only focus on numerical characteristics of UGC; research papers that investigate the antecedents and motives for UGC generation; and research papers in the fields of medicine, nutrition, food science, public administration, sports, etc., rather than e-commerce. Only papers with the aim of extracting valuable information and knowledge from UGC text using mainly text mining techniques for the purpose of business applications in the field of e-commerce were included. Review papers covering topics within the scope of this paper were included, for example, review papers about opinion mining from online consumer reviews, spam opinion detection, customer sentiment mining, etc.

As shown in Figure 1, the number of articles has increased significantly since 2015, which indicates that employing text mining techniques for business intelligence is increasingly attracting the interest of academics. We anticipate that more relevant studies will be published in the coming years.

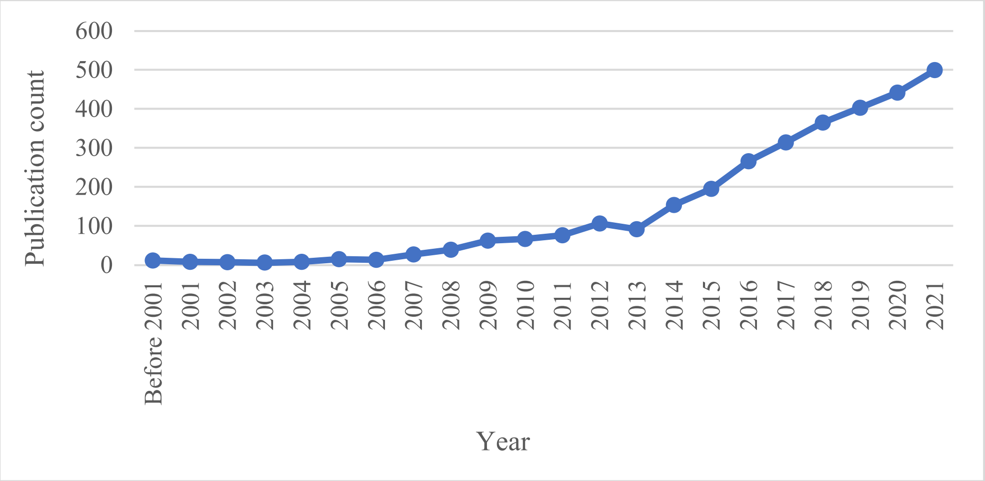

Figure 1. Publication timeline.

Tables 1 and 2 show the journals and conferences that have published a high number of articles on related topics and corresponding counts. The academic categories of these journals range from business and management to computer science and artificial intelligence. The former primarily concentrates on obtaining business insights, whereas the latter is more focused on improving text mining techniques by virtue of advanced methods generated in the area of computer science.

Table 1. List of journals and counts.

Sources	Count
EXPERT SYSTEMS WITH APPLICATIONS	57
DECISION SUPPORT SYSTEMS	56
IEEE ACCESS	51
arXiv	41
ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING	39
ELECTRONIC COMMERCE RESEARCH AND APPLICATIONS	31
SUSTAINABILITY	31
ELECTRONIC COMMERCE RESEARCH	28
COMPUTERS IN HUMAN BEHAVIOR	24
INTERNET RESEARCH	21
INDUSTRIAL MANAGEMENT & DATA SYSTEMS	19
INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS	17
INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT	17
KYBERNETES	17
SOFT COMPUTING	16
ASIA PACIFIC JOURNAL OF MARKETING AND LOGISTICS	15
JOURNAL OF THEORETICAL AND APPLIED ELECTRONIC COMMERCE RESEARCH	15
ONLINE INFORMATION REVIEW	15
INFORMATION SYSTEMS RESEARCH	14
JOURNAL OF BUSINESS RESEARCH	14
KNOWLEDGE-BASED SYSTEMS	14
BRITISH FOOD JOURNAL	13
INFORMATION SCIENCES	13
JOURNAL OF RETAILING AND CONSUMER SERVICES	13
INFORMATION PROCESSING & MANAGEMENT	12
MULTIMEDIA TOOLS AND APPLICATIONS	12
TELEMATICS AND INFORMATICS	12
APPLIED SCIENCES-BASEL	11
ELECTRONIC MARKETS	11
EUROPEAN JOURNAL OF MARKETING	11
INFORMATION & MANAGEMENT	11

Table 2. List of conferences and counts.

Sources	Count
ACM International Conference Proceeding Series	57
International Conference on Information Systems, ICIS	16
Journal of Physics: Conference Series	15
Proceedings of the Annual Hawaii International Conference on System Sciences	9
Proceedings of the International Conference on Electronic Business (ICEB)	9

An increasing number of journals encourages interdisciplinary studies based on methods, methodologies, and theories drawing from diverse disciplines. For example, as shown in Table 1, COMPUTERS IN HUMAN BEHAVIOR, which is generally categorized as PSYCHOLOGY; EXPERT SYSTEMS WITH APPLICATIONS, which traditionally aims covers topics of engineering and computer science; and ELECTRONIC MARKETS, which focuses on networked business, have been publishing research involving business applications of UGC text mining techniques. We expect that more cross-discipline studies will be published by journals whose scopes lie in business and computer sciences applications.

2.1.2. Thematic Analysis

Because the text mining of UGC is a cross-research field, a review of the research questions in the literature reveals that scholars have studied this field for a variety of research purposes, such as user satisfaction mining to improve recommendation accuracy, adaptive construction of product lexicons, review helpfulness identification, etc. In general, such studies generally involve semantic analysis, opinion mining, sentiment analysis, and other technologies to identify review characteristics, deeply understand user behavior patterns, and extract user needs with respect to product or service features to achieve user-centered product development and recommendation. Based on the keywords of the 3392 extracted literature publications, a thematic analysis was performed to identify the research contents and directions in this field. Table 3 shows that that sentiment analysis and user preference mining have received the most attention in recent years, representing hot spots in the field.

Table 3. Thematic analysis and share of search hits.

Research Topic	Research Items	Representative Papers	% of Hits
Semantic analysis	Feature extraction Topic mining Lexicon building Text representation Semantic understanding	[2,10–34]	16.63 (564/3392)
Sentiment analysis	Sentiment mining Opinion mining Sentiment classification	[5,10,13,35–54]	72.58 (2462/3392)
Text quality mining	Review helpfulness EWOM helpfulness Credibility Review ranking Review quality Spam detection Fake review dectection	[38,55–74]	10.08 (342/3392)
Consumer profiling	Preference analysis User requirements/needs/demands Requirement and expectation User satisfaction Personality	[2,5,6,12,13,27,75–92]	7.52 (255/3392)
Product design	Attribute performance Performance analysis Product/service development Product/service improvement Product/service attributes Product/service features Product/service quality	[4,86,93]	6.63 (225/3392)
Marketing and recommendation	Marketing strategies Recommendation systems Product recommendation Recommenders Product ranking Promotions	[69,80,94–99]	18.66 (633/3392)

Refining these studies will help to extract more valuable information from UGC to support production practices. According to the results of the thematic analysis of the literature, some of these studies focus on the design and optimization of text mining techniques, especially semantic mining and sentiment analysis. Some focus on the practical application of these techniques, such as review quality evaluation based on content characteristics, as well as consumer profiling, preference analysis, product design, and recommendation. Furthermore, some studies organically combine text mining technologies and applications of to solve specific practice problems. The main contents of this paper were determined according to the thematic analysis. Because these articles generally take UGC as the data source, we discuss the text mining of UGC from three perspectives: data input, text mining techniques, and their business applications.

2.2. Overview of the Research Framework

Based on our review of the literature, we found that the included studies generally followed a common process. Specifically, in the research field of UGC-based text mining, the data input is the content created directly by users and the additional information generated by their purchase behaviors. After data collection, effective data preprocessing is an indispensable step for subsequent analysis, as UGC in the form of text is unstructured [100]. In order to achieve the goals of text mining, such as extraction of user sentiment, a series of text mining models have been proposed, and various models and algorithms, such as statistical analysis, neural networks, and machine learning, have been applied for specific business applications, e.g., selection of high-quality UGC, consumer profiling, and making accurate product recommendations. Figure 2 shows the common framework of these studies. Next, we will carry out a detailed analysis of the three aspects of UGC text mining.

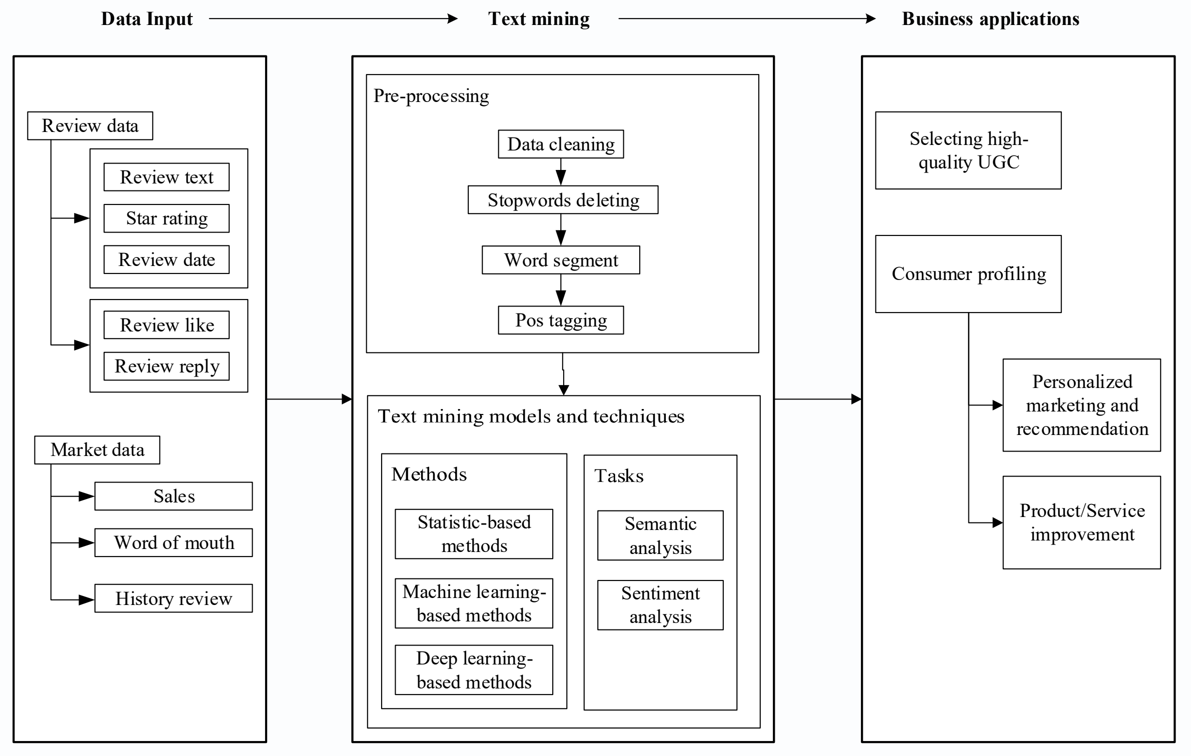

Figure 2. The general research framework.

2.2.1. Input Data

Generally speaking, text mining research based on UGC generally takes review data and market data as input. Review data are mainly related to users' online comments published after they consume a product or service, including star reviews, satisfaction level, review time, and review text [101]. Among them, review data in the form of text correspond to a variety of sources and types, such as reviews on online shopping platforms, user reviews on official brand websites or news websites, product-related user-generated content on social media platforms, etc. With the update of comment functions on e-commerce platforms, the likes and replies of other users to a review also attract attention and become part of the UGC data [55]. In addition, some studies comprehensively consider market data and historical data of users to expand review data, such as product sales, WOM, and historical user reviews [102]. Furthermore, market data often supplement review data to support such research. In general, product sales, word of mouth, and historical click or purchase data are common forms of market data.

2.2.2. Preprocessing

Unstructured UGC data contains a considerable amount useless information. In order to reduce noise in data analysis, effective data preprocessing has become a basic step in this field. Skoric et al. [103] investigated the role of text preprocessing in sentiment analysis of film reviews, proving that preprocessing can significantly improve the accuracy of sentiment feature extraction through proper feature representation. Generally, data cleaning (including deleting invalid content and stop words), word segmentation, word frequency statistics, and part-of-speech analysis are the main steps of text preprocessing [100]. First, in terms of UGC data cleaning, most scholars remove invalid comments (such as comments composed entirely of symbols, default comments, and irrelevant comments) and repetitive comments before conducting their analysis. In the subsequent text analysis process, stop words, such as "however", "and", and "is", that do not indicate specific semantics are deleted according to the existing stop-word lexicon [104]. Secondly, word segmentation and word frequency statistics are the premises of topic mining and sentiment analysis. English review texts are often processed by NLTK, SpaCy, and StanfordCoreNLP [35,56], whereas Jieba is the most widely used Chinese text preprocessing tool [75].

2.2.3. Text Mining Models and Techniques

Preprocessed UGC text is still unstructured, and the means by which to transform such text into a statistical and even computable structured form has attracted considerable attention.

Traditional work often lies on the statistical representation of items (e.g., words and sentences) in text [57], neglecting the contextual information of items. With the rapid development of machine learning and deep learning models, such models have been widely used to vectorize text to extract critical embedded features and realize text understanding tasks, such as similarity recognition, semantic analysis, and opinion mining [10]. For example, Rangnekar et al. [11] proposed a semi-supervised method combining word vector representation and K-means clustering to mine the features of user reviews from massive texts. In addition, some scholars use coding methods to structure terms in reviews according to specific procedures and analyze product or sentiment characteristics in a targeted manner [12,36].

3. UGC in E-Commerce: Sources and Types

In the Web 2.0 era, users have transformed from consumers of content to producers of personalized and individualized content. Users are allowed to directly publish content in any form, including text, images, audio, and video. In traditional e-commerce, UGC tends to refer to the concept of word-of-mouth (WOM) and online reviews on e-commerce websites, which constitute positive or negative evaluations and descriptions of a product or company generated and spread by potential, current, or former consumers [93]. Amazon, eBay, Mytheresa, Taobao, Jingdong, Airbnb, Yelp, and Tripadvisor are well-known e-commerce platforms that gather various kinds of products with a large number of users therefore targeting a large volume of UGC. Table 4 shows some mainstream e-commerce websites that are frequently used as sources of UGC. Abundant research focuses on mining the big data of review text on various e-commerce platforms for a direct and meaningful representation of consumers' experiences, attitudes, opinions, and sentiments. Besides review text, as shown in Table 5, data about ratings, sales, and interactions, as well as individual data and other supplementary data indirectly produced by consumer behaviors can provide a tremendous amount of information that reveals users' preferences, especially in the absence of purchases. Ghose et al. [105] combined an optimal stopping framework with an individual-level random utility choice model to analyze click behavior in conjunction with purchase choices under the social media context. Yuan, Xu, Li, and Lau [102] took into account both online reviews and commodity sales data, embedded consumers' daily emotions into the underlying thematic sentiment model, and realized the forecast of sales in the next period. Qiu and Cho [76] extracted product features from online reviews and constructed a user preference identification model based on historical user click data.

Table 4. Sources of UGC in practice.

Websites	Studies
Online shopping malls	
Amazon	[106,107]
JD	[5,77,106]
Tmall or Taobao	[106,108]
On-demand services	
Meituan	[109]
Dianping	[110,111]
Grubhub	[78]
Travel websites	
Tripadvisor	[79,112–115]
Ctrip	[116,117]
Review websites	
Yelp	[56,106,118–121]
Trustpilot	[121]
Social media platforms	
Twitter	[122]
Facebook	[123]

Table 5. Additional data types.

Additional Data	Description	References
Ratings	Star ratings or scores of each review record.	[13,37,80]
Sales data	Market-level data of a product/brand concerning sales	[81,124]
Interaction data	The comments, replies, favorites, likes, visits, and shares of each UGC item	[38,56,58,59]
Individual data	Demographic characteristics, personalized preference characteristics, individual tags, and historical behavior data	[76,125,126]
Supplementary data	For example, opinions from experts and surveys	[57,82]

With the rapid development of Web 2.0 and the emergence of social commerce, the concept of UGC expanded. On the one hand, social commerce is often considered a subset of e-commerce [9] because it realizes the same purpose of commercial activities as e-commerce through more interactive means. Web 2.0 provides the means to harness business intelligence by allowing a wide range of activities, such as commenting on a product, commenting on a comment, leaving questions, replying, visiting, discussing, "liking", and sharing [127]. Some scholars consider these interactive contents as the expansion of review data to enrich research on online review mining. For example, Li, Huang, Tan, and Wei [59] took text features as an independent variable and the helpfulness represented by likes, favorites, and replies as a dependent variable to analyze the mechanism by which users perceive the helpfulness of reviews.

In essence, social commerce does not merely comprise social functions; it can considerably lead business evolution from a product-centered perspective to a consumer-oriented perspective [127]. Traditionally, consumers make choices depending on recommended results of search engines. Therefore, a stream of research examines how a product can outperform in the search and decision stage of consumers [105]. However, abundant UGC can convey more consumer information and generate a deep understanding of consumers, leading to consumer-oriented marketing and product improvement.

Online users generate a large volume of content at an overwhelming speed. Although UGC is deemed beneficial, not all UGC is valuable all purposes [60]. Therefore, excessive and low-value user-generated text contradicts the requirement to understand consumers deeply and draw knowledge embedded in a specific application. To effectively and artificially mine textual UGC big data, variant text mining techniques have been used and combined both in academia and the practice of e-commerce. In the next section, we summarize mainstream innovative research on the application of textual UGC in the field of e-commerce employing text mining.

4. Techniques for Text Data Mining

Text mining, also known as text data mining, is the method of automatically obtaining reliable and meaningful information patterns from the unstructured form of data [128]. Given that text is unstructured and is difficult to automatically interpret and understand, the primary task of text mining is to structure the text content, including text parsing, vectorization, linguistic feature extraction, topic modeling, and sentiment analysis. To fulfill these tasks, text mining techniques represent a truly interdisciplinary method that draws on machine learning, statistics, and computational linguistics [129]. Hereafter, we provide a short summary of frequently used techniques of text mining and exemplify typical methods, such as algorithms and models, as well as corresponding implementations in practice.

4.1. Semantic Analysis

The aim of semantic analysis is to evaluate and represent written text and analyze the meaning of sentences with interpretations similar to those of human beings [14]. Text representation and semantic feature extraction are the basis of further understanding, and text semantic understanding is the up-level task of text semantic analysis.

4.1.1. Text Representation

In the process of text semantic analysis, the text should first be represented as vectors in a high-dimensional space. Text representation techniques include statistics-based methods and deep-learning-based methods.

1. **Statistics-based methods**

Traditional research involves attempts to vectorize text by counting the frequency of words. The bag-of-words model uses the frequency of an unordered set of words to express a sentence or a document. It views a text as simply a set of words, neglecting grammar and the order of words in the text.

Based on the bag-of-word model, the n-gram model considers a contiguous sequence of n items as a gram and employs frequency statistics for each gram to form the vector space of a text. N-gram assumes Markov properties, that is, the nth item can be inferred from the first n-1 items. The probability of a sequence as a whole is the product of the probabilities of each item's occurrences. Items in the n-gram model are generally words or in some cases, phonemes, syllables, letters, or sentences. For example, employing the bigram model, Meng et al. [15] proposed a bigram sentence language model to calculate the occurrence probability of a given sentence. Based on probability, the information entropy and perplexity of the sentence can be calculated to represent the readability of a sentence.

Term frequency-inverse document frequency (TF-IDF) is a numerical statistics approach focusing on key terms in a document. The importance of a term increases with its occurrence in a document and decreases with its frequency in the whole corpus. Burtch et al. [16] measured the cosine distance between two document vectors represented by a TF-IDF model. According to the similarity, the novelty of a given consumer review was further calculated in terms of word importance and word frequency.

2. Neural-network-based methods

Neural network methods have provided novel solutions for text representation and achieved remarkable results compared to statistics-based methods [17]. Neural network models can not only learn the words themselves but can also capture the location and context of a word.

Word2vec, the most commonly used word representation method, applies a continuous bag of words (CBOW) and the skip-gram algorithm and introduces the Hoffman tree and negative sampling to improve both the quality of the vectors and the training speed [18]. The skip-gram model is able to predict the context of a given word, whereas the CBOW model predicts words according to the context. Based on word2vec, doc2vec, an unsupervised framework, was proposed to learn continuous distributed vector representations for texts ranging from sentences to documents [19]. For instance, Burtch, He, Hong, and Lee [16] finetuned a pretrained doc2vec model based on their dataset to obtain a vector representation of heterogeneity across users posts, which was then applied to measure the creativity of a given post from the document.

Besides these well-established models, in recent research, feed-forward neural networks (FNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), and recursive neural networks have been commonly used, modified, and combined to vectorize words, sentences, or text for further semantic understanding [20].

3. Transformer-based model

The transformer-based model outperforms in-text semantic feature extraction. Bidirectional encoder representations from transformers (BERT) is a well-known and well-established transformer-based structure. BERT is a deep bidirectional language representation that can integrate left and right contextual information. BERT can be pretrained and finetuned for a wide range of tasks and has achieved state-of-the-art results on some tasks, including text classification, semantic understanding, etc. [21]. Recently, improved language models based on BERT have been gradually emerging, attaining considerable robustness and effectiveness in specific tasks.

However, as a transformer-based model, BERT relies heavily on the knowledge derived from the experience learned from a large corpus [22]. In addition, BERT models require considerable computing resources in the training stage. BERT appears to contain appropriate mechanisms for learning universal linguistic representations that are task-independent [22]. Therefore, publicizing BERT models that have been pretrained based on various types of corpora can provide pretrained models to help other researchers finetune their models according to specific tasks, contributing to efficiency improvement of research and business practice.

4.1.2. Semantic Understanding

To capture the semantic sense of a document, text should be interpreted in a way that human beings can understand. The topic model, which is the most commonly used unsupervised method, enables the discovery of human-readable "topics" that occur in a set of documents, helping to uncover hidden semantic structures.

Latent semantic analysis (LSA), latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), and BERTopic are the most widely used topic models. Specifically, LSA, also called latent semantic indexing (LSI), is a procedure for examining the associations among a massive set of documents and the words/phrases included by generating ideas belonging to articles and terms. LSA employs singular value decomposition (SVD) to decompose and reduce the term–document matrix to obtain a more present semantic structure of the document [23]. LDA assumes that a document is a mixture of underlying topics and that a topic is a collection of terms. For a given statistical representation, the LDA model provides the probability distribution of the topic for each document in the set to summarize a document at the topic level [2]. In contrast to LDA and LSA, NMF is a decompositional, non-probabilistic algorithm using matrix factorization. It works on TF-IDF-transformed data by breaking down a matrix into two secondary matrices [24]. BERTopic is a topic-modeling technique that leverages transformers and c-TF-IDF to form dense clusters, allowing for explicable topics while keeping important words within the topic descriptions. BERTopic supports guided, (semi-)supervised, and dynamic topic modeling [25]. Extant research demonstrates that BERTopic can provide more clear-cut topics and generate more novel insights than LSA and LDA with respect to short texts [24]. Filieri et al. [26] employed LSA to separately extract critical service features from positive and negative online reviews. Guo, Barnes, and Jia [2] identified the key dimensions of consumer service voiced by hotel visitors using LDA.

Furthermore, methods such as the conditional random field model, frequency-based feature extraction, and rule-based feature extraction are often applied. For sentiment analysis of UGC, Xia, Yang, Pan, Zhang, and An [10] used the conditional random field algorithm to extract sentiment features from review segments and applied an SVM-based classifier to obtain asymmetrically weighted sentiment values of all feature words. Cai et al. [27] extracted product features from reviews by machine learning, taking the frequency of user description of this feature as the attention index, and weighted the sentiment value to obtain a ranking of user requirements. Araie and Takase [28] designed a feature word extraction rule unit and a rule template based on the word form, part of speech, dependency relation, control word, and emotional description and realized the effective extraction of product feature words based on conditional random fields.

In the field of UGC mining, the topic model is also known as “aspect extraction”, which automatically obtains the key elements of a text from a large corpus of UGC [29,30], including product characteristics [31] and specific themes [32]. The main description object of reviews is generally the product itself, and the functions and attributes of the product constitute the product characteristics [33]. Park and Kim [12] focused on the critical components of the product in practical application, detected phrases from online reviews and represented them with word vectors, and finally determined the representative product characteristics through cluster analysis. These methods have been commonly used in previous research. In addition, some studies have mined UGC for specific topics, such as users’ personal characteristics. Kim et al. [34] designed a dual process model to extract argument quality, review valence, review helpfulness, information laterality, source credibility, and reviewer recommendation, realizing the mining of multiple specific topics.

Features explicitly mentioned by the user, i.e., explicit features, are easy to extract from UGC. However, implicit features are much more difficult to identify than explicit features, including (1) features that are ignored in user reviews and (2) features that are mentioned by users in reviews but that cannot be directly recognized by machines due to sentence ambiguity or unclear pronoun reference [11,33]. To mine user preferences with respect to product attributes embedded in incomplete online comments, Li, Zhang, Li, and Yu [13] proposed a comment extension mining model to patch online reviews based on semantic similarity and emotional resemblance. However, little research has been published considering attention to implicit feature extraction, which has considerable potential as a future research direction.

4.2. Opinion Mining and Sentiment Analysis

Sentiment analysis and opinion mining, which are often used interchangeably, focus on systematically identifying, extracting, and quantifying people’s affective states and subjective information from written text [39]. User sentiment in UGC comprises attitudes toward products or services and indicates the degree to which requirements are satisfied [40,41]. Sentiment analysis in the field of text mining is based on sentiment words in text, which are the fundamental basis for judging users’ attitudes [42,43] and analyzing the voice of the consumer for applications that range from consumer research to marketing and product development. Research on sentiment analysis can be divided into coarse-grained (document-level and sentence-level) sentiment analysis [44] and fine-grained (feature-level) sentiment analysis [10,45]. In UGC analysis, the former refers to mining the emotional attitude of the whole UGC text, whereas the latter identifies a user's emotion toward product features or target topics in a more granular manner [36,46]. Figure 3 illustrates the common flow of sentiment analysis for UGC.

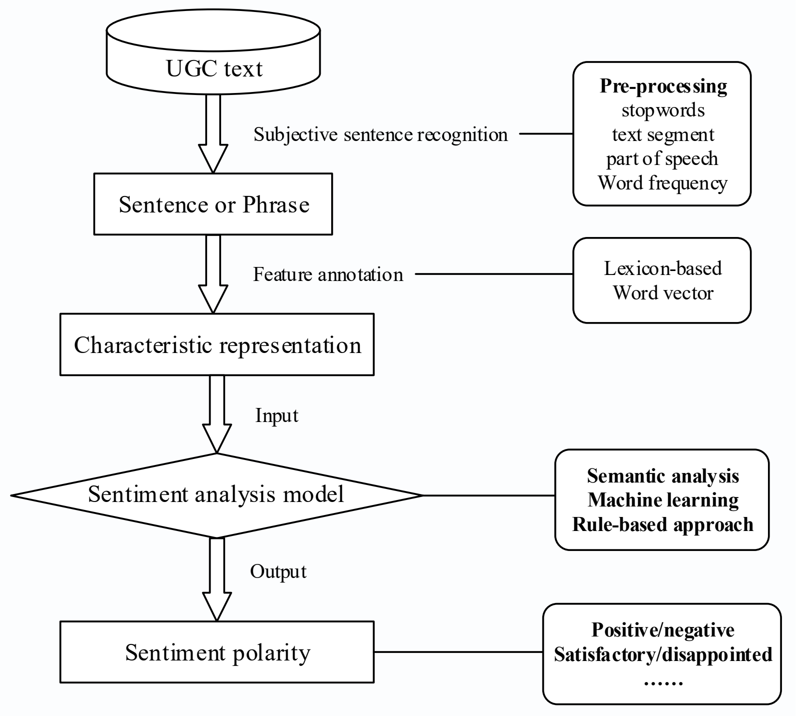

Figure 3. Sentiment analysis process of UGC.

Sentiment analysis techniques can be divided into three categories. A sentiment lexicon is a tool commonly used for sentiment analysis based on semantic relations [38,47]. The lexicon-based approach makes use of a sentiment lexicon to compute the global sentiment polarity of a text document, depending on the polarity of the individual words incorporated in the document [48]. Commonly used sentiment lexicons includes SentiWordNet (English), the National Taiwan University Sentiment Dictionary (NTUSD) (Chinese), HowNet (Chinese), and NLTK [35] (English). Some scholars proposed a fuzzy algorithm based on a sentiment lexicon to identify the positive, neutral, or negative sentiment orientation of alternative products in each review [47]. On the one hand, this lexicon-based method fails to consider the overall sentiment of the context and is limited by the number of words, failing to meet the needs of UGC sentiment analysis. On the other hand, given that the existing lexicon may be not suitable for sentiment phases in a specific context, some researchers have built a sentiment lexicons based on a specific corpus. To precisely understand user requirements embedded in reviews on cell phones, Qi, Zhang, Jeon, and Zhou [5] built a sentiment lexicon including 554 positive words and 355 negative words.

The machine learning-based approach, also known as the corpus-based method, treats sentiment analysis as a binary (positive or negative) or multiclass classification problem [49]. The aim of this approach is to classify and predict the sentimental polarity of a given text based on a machine learning (ML) model constructed and trained by input text with sentimental labels. In ML, naïve Bayes (NB) achieves excellent performance in text classification, and support vector machine (SVM) has advantages over small samples [37]. SVM is a learning system that employs a linear function hypothesis in a high-dimensional feature space. Based on Bayes' theorem, naïve Bayes assumes that the probability of occurrence of each feature is independent and predicts the category of each individual with its features using the prior probability of each category and the conditional probability of occurrence of specific features in each category. Snownlp, a widely used Chinese text analyzer implemented in Python, adopts a naïve Bayes algorithm to train review data with the positive or negative label of e-commerce products and services to evaluate the sentiment score. For example, Li et al. [50] employed Snownlp to infer reviews collected from Autohome, China's largest online car community. In order to study the comprehensive effect of machine learning classifiers and sampling methods on sentiment classification, Vinodhini and Chandrasekaran [51] modified the SVM-based ensemble algorithm by combining oversampling and undersampling to improve the prediction performance. Su and Shen [52] established product and sentiment features based on expert knowledge and fuzzy mathematics and constructed a convolution attention long short-term memory (CA-LSTM) model, fully considering the long-term dependence between features to achieve sentiment classification. These methods can be divided into supervised models and unsupervised models [53]; supervised models represent the mainstream of this stream of research.

Some scholars have attempted to use deep learning methods for sentiment analysis [35,52]. The most commonly used models are CNN, RNN, LSTM, etc., although research on sentiment analysis using deep learning is still insufficient. BERT, a well-established structure based on transformers, was employed to build a pretrained sentiment classification model. For example, Munikar et al. [54] employed BERT to a fine-grained sentiment classification task. Experimental results showed that the proposed model achieved excellent compared to other popular models, including RNN, LSTM, and CNN, without sophisticated architecture.

Researchers have also attempted to analyze sentiment by combining several methods. Li, Zhang, Li, and Yu [13] roughly identified the explicit emotions of users based on a sentiment lexicon and designed a deep neural network based on semantic similarity and sentiment similarity to identify the implicit emotions not mentioned in reviews. Such a method combines a sentiment lexicon and machine learning to ensure the accuracy of sentiment analysis and improve mining efficiency, representing a promising research direction in this field.

However, these studies mostly rely on a large number of labeled samples, which are time-consuming to collect, requiring considerable research effort. It remains challenging to obtain precise and robust sentiment analysis results based on a small, labeled dataset. In addition, a well-established, pretrained model in a specific domain has the potential to improve the precision of sentiment analysis in a cost-effective manner.

5. Applications in E-Commerce

To overcome the problem of information overload, a large volume of UGC has to be identified to select high-quality, credible, and useful content for various application purposes. Through semantic and sentiment analysis, UGC can be broadly used to profile consumers in various aspects, based on which businesses can improve marketing and products/services in a consumer-oriented manner. To this end, domain knowledge and theories from various disciplines have been combined with text mining techniques.

5.1. Quality Evaluation of UGC Text

Hu, Bose, Koh, and Liu [57] pointed out that the efficiency and accuracy of UGC mining are based on high-quality reviews, i.e., reviews should be complete, clear, emotional, and credible. Text quality determines whether such data can be applied in practice [61]. Existing studies have adopted helpfulness as a critical indicator to measure the quality of UGC. Review ranking and spam UGC detection rely on content quality evaluation.

5.1.1. Review Helpfulness

Although UGC texts are deemed useful, not all of them actually are. Excessive UGC and low value density lead to information overload [60] and confuse both consumers and businesses. Hence, distinguishing helpful reviews has become an arduous task. Review helpfulness has been well-studied in this field, referring to the ability of a review to provide useful information to the reader and assist them in making decisions [38,62]. For users, review helpfulness helps to eliminate uncertainty with respect to purchases [63]. Given that overload can strain consumers' cognitive capabilities and increase the difficulty of information processing, helpfulness can be used by users as a quality indicator. As shown in Figure 4, some research has focused on determining the critical factors that affect review helpfulness, generally encompassing textual characteristics extracted from review content, in addition to reviewer- and source-related factors [64].

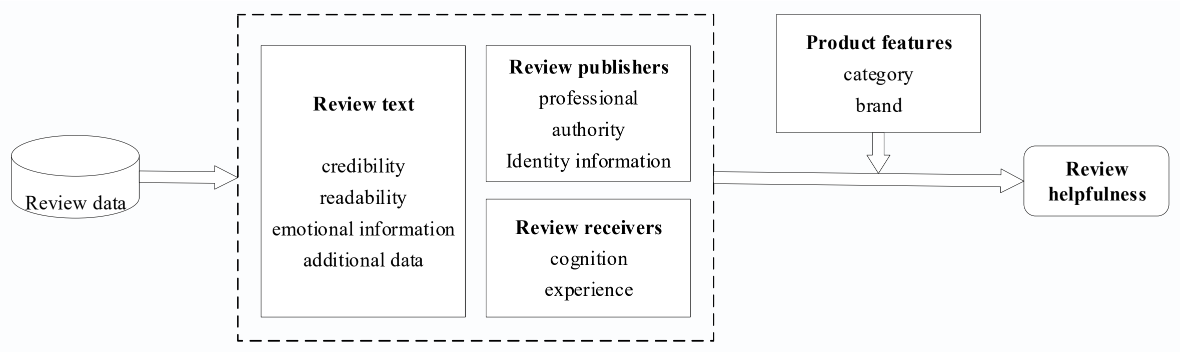

Figure 4. Research framework of online review helpfulness.

As the most important type of UGC, review content has received the most research attention. Li and Zhan [65] explored how language style, organizational structure, and other content characteristics affect the perceived usefulness of reviews. Similarly, Kang and Zhou [66] investigated the breadth, depth, and redundancy of online reviews by introducing semantic hierarchy and proposed a series of indicators, such as the uncertainty of review content, information quality, signaling, and coding variability, to reveal the influencing factors of review readability. Among these factors, review credibility has attracted the most attention. For example, Nilsson and Hultin [67] designed a mainstream feature viewpoint pair based on big data and developed review credibility measurement rules based on user satisfaction levels with respect to these features. Emotional information is also an important factor in evaluating review quality. Lee et al. [68] selected online movie reviews as data, presented a mining method for emotional entropy of review texts, and quantitatively measured the relationship between emotion and usefulness. Negative emotion significantly enhances the perceived usefulness of reviews. In addition, comments, retweets, and likes of online reviews are considered in relevant fields as mutually causal factors [55].

Secondly, review publishers influence the credibility and usefulness of review quality evaluation. To study the influence of a reviewer's credibility on sales, Banerjee, Bhattacharyya, and Bose [56] proposed specific characteristics of a reviewer's credibility according to the source credibility theory, involving professionalism, authority, enthusiasm, engagement, experience, reputation, ability, social ability, etc. In addition, some scholars have included the characteristics of review receivers in their studies. For example, Filieri, Hofacker, and Alguezaui [58] considered the potential moderating effect of review receivers' participation and investigated the influence paths of central and peripheral cues, such as relevance, authenticity, source credibility, and ranking score based on the refined possibility model, further describing the mechanism of perceived review helpfulness.

5.1.2. Review Ranking

The aim of review ranking is to screen and prioritize high-quality reviews to ensure that users have easy access to information that is beneficial to their purchase decisions [69]. The number of online reviews has exploded in recent years, and many e-commerce platforms provide a review-ranking function, using consumer feedback to determine the quality of each review and ranking them according to their quality [65,67]. Initial review ranking is based only on the recognition degree of other users (for example, the Amazon platform includes a "like" function for reviews); however, this method is based on the premise that a large number of users have read the review in question, resulting in the problem of unbalanced voting (reviews posted earlier will be seen more often, and those with more votes will receive more additional votes) [58]. The application of text mining techniques elicits new ideas with respect to this scientific problem by automatically extracting semantic quality [59] and sentiment strength [68] from online reviews and even considering the information about the reviewers to determine the usefulness and credibility of the reviews [56].

Based on a review of relevant literature, we found that existing research on review ranking can be divided into two categories: (1) generation of a non-personalized review ranking for all consumers without differentiation [38,65,70] and (2) personalized review ranking for individual consumers [69,71]. The former has been more widely studied than the latter, and scholars have proposed review-ranking models based on the relevant indicators involved in text quality evaluation. For example, Li, Huang, Tan, and Wei [59] proposed a product-review-ranking method based on sentiment analysis technology and intuitionistic fuzzy set theory. The authors conceptualized review helpfulness as a second-order formative structure, including perceived source credibility, perceived content diagnosability, and perceived alternative expression. Suryadi and Kim [70] used a combination of word embedding and the X-means clustering method to identify product feature words and sentiment words, according to which the relationship between online review rankings and sales was characterized. Using business reviews on Yelp.com, Banerjee, Bhattacharyya, and Bose [56] first assessed the influence of reviewer credibility (including enthusiasm, participation, experience, reputation, ability, and social ability) on business sponsorship to help enterprises judge and sort trustworthy reviewers.

Moreover, some scholars have attempted to personalized review rankings. For example, Cezar and Ogut [69] explored the influence of review ratings (location rating and service rating), product recommendations, and review display order on user conversion rate, taking the individual information of review recipients into consideration and developing a conversion rate model using the Bernoulli log-likelihood function and a parametric regression model based on beta distribution. Dash, Zhang, and Zhou [71] suggested that user preferences may differ in terms of product features and proposed a personalized review-ranking framework based on product features; the latent category regression model was used to predict the review helpfulness for individual consumers according to their preferences for product features.

5.1.3. Spam UGC Detection

Although UGC is an abundant source of consumer information, its validity is often questioned. It has been demonstrated that most consumers refer to and are prone to make decisions based on reviews or content generated by consumers rather than firms and marketers. By taking the advantage of the large number of users of electronic websites and social media, UGC can be propagated on a large scale to reach more targets in a timely manner [72]. The influential power of eWOM encourages the emergence and spread of fake reviews. Spam UGC could result in the loss of credibility and lead to consumer confusion about a product or mislead businesses and cause a misunderstanding of consumers [73]. Therefore, a stream of research on fake UGC detection has proposed supervised and semi-supervised algorithms, which rely heavily on textual and behavioral features extracted from UGC [72].

Some typical clues embedded in reviews have been suggested to identify fake content. Such indicators include scarce information about the reviewer, similarity with other content, short reviews, a focus on personal information, and frequent use of positive and negative words [73]. Kauffmann et al. proposed that sentiment analysis and neural networks can be used to detect spam reviews by exploiting product-related review features. To overcome the problem that the supervised algorithms require a labeled dataset, Saumya and Singh [74] proposed an unsupervised learning model, LSTM-autoencoder, which was trained to learn the patterns and sequences of words and sentences in real review from the review's textual details without labels.

However, the above studies fail to consider the relevant characteristics of UGC quality from the perspective of products/services, resulting in analysis results that are not fully applicable to specific product/service involvement. Furthermore, most extant research has adopted a fixed and single method to evaluate UGC quality, neglecting the differences between consumers. In particular, review ranking and spam detection based on user preferences and individual information remain to be further expanded.

5.2. Consumer Profiling

Consumer profiling, also called user labeling and user modeling, refers to the definition of users through a series of concise and differentiated attributes, descriptions, and labels. Generally, consumer profiles contain critical information that is used to identify an individual, including demographic and individual characteristics, such as age, name, knowledge, or expertise, as well as underlying characteristics, such as the user's preferences, interests, and behavior [83]. Previous research on consumer profiling has focused on the use of databases that contain digital traces of users, such as user behavior logs, page click history, and product transaction records. With the aid of massive UGC and the development of text analysis, users can be understood and profiled based on characteristics embedded in text, which has become an important research direction in the field of business intelligence [84]. For this purpose, text mining is necessary in this field. In the following sections, we provide an overview of the streams subject to consumer profiling, mainly relying on text mining.

5.2.1. User Requirements and Preferences

In the demand-driven economic era, comprehensively providing unique products to meet the requirements of consumers has become the key to success for enterprises. User preference analysis is a prerequisite for meeting user requirements. Preference is defined as the degree of interest and importance that users exhibit for given categories and is the comprehensive result of users' internal weighing of a product or a service based on their perceptions [80,82]. In recent years, a large and growing body of literature has investigated the extraction of user requirements and preferences from UGC [85]. Compared with consumer-survey-based analysis, UGC enables businesses to gain demand-side knowledge that is less biased and more valuable [86]. Through semantic and sentiment analysis, some researchers have attempted to infer user requirements based on factors and opinions extracted from text. For example, Kim and Na [87] examined the possibility of distinctive consumer requirements and awareness by analyzing reviews on cycling pants using text mining in online shopping malls. Based on sentiments extracted from reviews, Sun, Guo, Shao, and Rong [85] proposed a dynamic mining method to mine the change in user requirements and manufacturers' opinions over time.

Extant research has incorporated text mining techniques and theory in other disciplines to identify user preferences based on UGC. For instance, KANO, a classical preference model, has been widely used as an effective tool to portray user preferences [75,88]. Hou et al. [89] used a rule-based NLP approach to automatically identify and construct user preference features in product reviews, and inspired by the KANO model, they classified preferences into five categories according to their importance. Qi, Zhang, Jeon, and Zhou [5] proposed attribute preference determination rules based on KANO theory using online reviews as a source of data.

Considering individual variability, consumers interests differ depending on the user; thus, user preferences also exhibit heterogeneity [6,13]. Therefore, user preference identification is often built based on effective market segmentation. In general, segmenting consumer groups based on the variability of user preferences ensures that a given class of users is relatively homogenous [81]. For example, Li, Liu, Lu, Zhang, Li, and Yu [75] developed a template-matching deep mining method to narrow down the user preferences among groups through consumer segmentation and to determine user preferences for product attributes based on the emotional polarity of each consumer group, summarizing the common and personalized preferences of each group. Francesco and Roberta [90] constructed traveler profiles by measuring frequencies of hotel attributes in each review using text-linking analysis. The results suggested that there are significant cross-country differences in terms of how travelers perceive and emphasize several hotel attributes.

With respect to preference classification methods, clustering techniques based on the K-means algorithm, association rules, and content-based and collaborative filtering techniques are generally used. For example, Cai, Tan, Ge, Dou, Huang, and Du [27] extracted requirement characteristics from online reviews through the machine learning method and applied hierarchical clustering models to identify the requirements and preferences of different users. Park and Kim [12] proposed a new method to extract subfeatures from online data based on association rules and tested the proposed method on smartphone review data. To identify user preferences from the perspective of product improvement, Li, Zhang, Li, and Yu [13] proposed an extended review-mining model based on semantic and sentiment similarity for differentiated product improvement.

In conclusion, UGC text mining techniques have been applied extensively to identify user preferences, and the heterogeneous preferences and personalized requirements of consumers have received increasing attention from scholars in recent years. UGC provides new data sources for mining of individual consumer information [76], representing a prospective direction for future research.

5.2.2. Consumer Satisfaction

Review content enables the examination of consumers' online review-writing behavior, as it provides details about what consumers emphasize, how they think, and how satisfaction forms [79]. Compared with numerical or star ratings, text can convey consumer satisfaction in a finer-grained and more precise manner. Research on consumer satisfaction in this field focuses on extraction of critical attributes involved in the considerable amount of text data available online using text mining techniques and depicting the formation and capturing the determinants of satisfaction utilizing econometric models or machine learning models [91].

Hong, Zheng, Wu, and Pu [77] employed a CNN text mining model and correlation analysis to compare the significance of various dimensions of fresh e-commerce logistics services involved in consumer reviews. The results confirmed that convenience, communication, reliability, and responsiveness had a significant impact on consumer satisfaction, whereas integrity had none. Chatterjee, Goyal, Prakash, and Sharma [91] used synthetic text mining, machine learning, and an econometric model to determine the core and augmented service aspects and important emotions that are reflective and helpful in predicting consumer satisfaction. Using text mining approaches, Xu [78] examined the critical topics consumers commented on and identified determinants of satisfaction through consumer online reviews in the context of on-demand food service.

Researchers have paid attention to the heterogeneity of satisfaction formation across segments. Bi, Liu, Fan, and Zhang [79] explored the asymmetric effects of attribute performance on consumer satisfaction for different market segments, including various different types of hotels, various types of tourists, and tourists from different regions. Similarly, Guo, Barnes, and Jia [2] identified the key dimensions of consumer service voiced by hotel visitors using LDA and identified differences across demographic segments.

5.2.3. Consumer Personality

Personality plays a central role in describing a person and influences an individual's behavior in various contexts. In addition to psychological methods, Xia Liu et al. [92] proposed a framework to infer a reviewer's personality traits by synergistically using personality theories and NLP algorithms to convert the available information on a review platform.

Although the heterogeneity of user preferences has received some attention, research on consumer profiling is still at a disadvantage. In terms of depicting user requirements and preferences, UGC texts cannot provide comprehensive information about consumers. Social media connects users and allows for interactive activities among them and even across platforms. In the era of social commerce, intelligent business requires a consumer-oriented perspective by synthetically collecting and analyzing data of a target consumer from multiple sources to develop consumer profiles.

5.3. Product/Service Evaluation and Enhancement

By extracting valuable information embedded in UGC, researchers and businesses can obtain a deeper understanding of user requirements, what users prefer, how they make decisions, and what motivates their satisfaction; consequently, innovation or improvement in production and marketing can be implemented in a consumer-oriented manner. In contrast to conventional consumer surveys, a wide body of literature on consumer-driven innovation emphasizes the important strategic value of absorbing consumer demand-side knowledge and capturing consumer-oriented design elements to provide complementary effects for service design [86]. Extant research has combined domain knowledge with text mining methods to realize product improvement. For example, using the human associative memory model as a theoretical framework, Gensler, Völckner, Egger, Fischbach, and Schoder [4] suggested a method to transform online product reviews into valuable information about brand images from the consumer perspective by combining text mining and network analysis. The approach aided businesses in efficiently tracking and identifying brand image weaknesses. To derive insights into the development and differentiation of cross-border logistics services (CBLS), Hsiao et al. [130] applied text mining to identify service elements and Kansei words in online reviews. Furthermore, their study exemplifies the integration of a traditional engineering approach (Kansei engineering) with UGC analysis to obtain ideas for service design.

5.4. Personalized Marketing and Recommendation

Nowadays, consumers are overwhelmed by a large number of product recommendations and content sharing. Effective marketing and recommendation require sufficient data and analysis to precisely identify targets, create attractive content, derive consumer engagement, and improve user experience [94].

Given that UGC contains abundant information about users, it represents an important source of data for research in the area of recommendation systems [95]. In recent research, personalized recommendations have widely incorporated the characteristics of users, such as user knowledge, expertise, product involvement, cultural background, and personal characteristics [69]. Some studies report the categorization of users into groups based on their interests, requirements, and preferences, relying on the importance of products and information for each user group to determine the corresponding recommendation items. For example, Liu, He, Wang, Song, and Du [80] proposed a new recommendation algorithm based on online review analysis that considered explicit ratings and implicit opinions to identify user preferences and construct a restaurant recommendation system for differentiated users. Deng, Gao, and Vuppalapati [94] developed a big data mobile marketing analytics and advertising recommendation framework to support both offline and online advertising operations based on collected data on mobile users' profiles, access behaviors, and mobility patterns.

The mainstream classification of existing recommendations is content-based and collaborative filtering-based recommendations [96]. Figure 5 shows the process of recommendation based on UGC mining. The content-based recommendation approach is used to predict the preference for another object based on the attribute similarity and relevance between the new object and reviewed items. For example, Forhad et al. [97] proposed an efficient hotel recommendation based on hotel attribute analysis by filtering and analyzing heterogeneous data from multiple sources, such as hotel reviews and word-of-mouth systems. Collaborative filtering-based recommendations obtain product recommendation lists by identifying a set of users who are similar to the current user. Based on this technique and online reviews, Wang and Wang [98] mined user opinions and mapped them to multicategory and single-category recommendations.

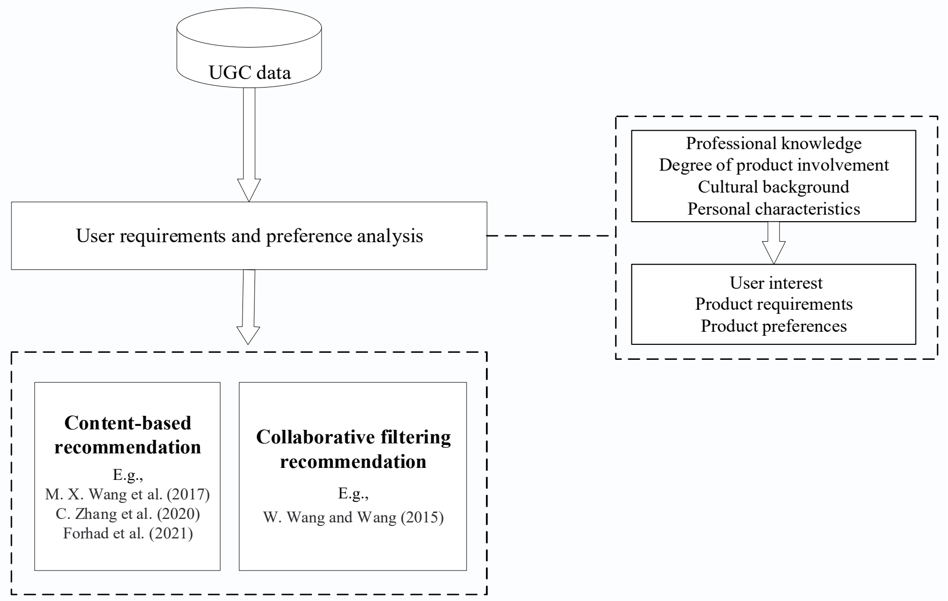

Figure 5. Requirement identification and recommendation system [97,98,110,131].

However, the pursuit of extreme personalization of recommendation and marketing is prone to cause “information cocoon” problems. When technology gradually matures to capture consumers’ attention based on their UGC and other related behavior (e.g., purchases, clicks, and views), personalized information that is highly matched with user interests and values is recommended, resulting in repeated recommendation of massively similar information [99], which is less beneficial for identifying latent or new interests and encouraging potential recommendation. Therefore, overcoming the extremely personalized marketing and solving the problem of “information cocoons” represents a prospective research topic.

6. Discussion

6.1. Text Mining Technologies

In the field of semantic mining, researchers have designed multiple feature extraction approaches to automatically capture key elements in text from a large quantity of UGC [29,30]. Among them, the mining of explicit features has been well-studied; however, the mining of features that are not explicitly mentioned in UGC is still challenging, and fine-grained and invisible feature extraction is rarely addressed [132]. A valuable future research direction is to identify the implicit features of UGC more precisely by considering contextual textual information (e.g., overall sentiment attitude and description objects) as well as individual information about the publisher (e.g., emotional state, cultural background, personality, etc.).

As the most popular method in the field of text mining, sentiment analysis has become the basis for application research [36,46]. Scholars in the field of UGC mining have successfully developed sentiment analysis to a fine-grained level, i.e., to identify users’ sentiments with respect to detailed product features or target topics [10,45]. Lexicon-based methods and machine learning techniques have become the mainstream for sentiment analysis. The former has the advantage of accurate sentiment recognition but excessively relies on sentiment lexicons, making it unsuitable for mining reviews related to emerging products or brands. The latter commonly applies machine learning algorithms, such as SVM [51] and neural networks [52]. Although it alleviates the dependence on the lexicon to some extent, supervised machine learning requires a large number of manually labeled samples. In addition, unsupervised machine learning methods and pretrained models are still in their infancy and need to be explored in depth.

6.2. Business Applications

Regarding the application of text mining techniques in e-commerce practice, we discuss the current status of research on UGC quality evaluation, consumer profiling, product/service enhancement, and marketing.

First, text mining supports the automatic extraction of textual features of UGC, as well as features of review publishers and review recipients, aiding review ranking decisions and spam UGC detection through metrics such as helpfulness and credibility [65,67]. Currently, undifferentiated review ranking for all consumers is the mainstream of this research [38,65,70], whereas generation of personalized review ranking for individual and differentiated consumers has rarely been addressed. Way to differentiate the presentation of reviews based on user preferences and individual information is a potential direction for future research in this field.

Furthermore, profiling of various aspects of consumers based on UGC, such as user interests, requirements, and preferences, is a considerable problem to be addressed. Considering the heterogeneity of consumers, researchers have attempted to segment consumers based according to various theories to extract the personality and commonality of consumers expressed in UGC. However, UGC text is insufficient to comprehensively understand consumers, and it would be a valuable research direction to portray consumers based on data about consumer online behavior from multiple perspectives sources.

Third, in the application of product improvement, although some classical preference identification models and product improvement models are available, providing a theoretical basis for this type of research, whether relevant features can be extracted from UGC to support the quantitative application of these theoretical models requires further in-depth exploration. In addition, UGC can support personalized consumer profiling [76], and the possibility of personalized product improvement or even product customization represents a valuable direction for future research in this field.

Fourth, UGC provides abundant information for personalized marketing and recommendations, and a series of studies have been conducted to generate and deliver recommendation lists that match the interests of users [125,131]. Although content-based recommendations and collaborative filtering-based recommendations have been extensively studied [96], the possibility of combining these methods based on refined UGC feature mining methods to recommend products to users more efficiently represents a prospective direction for future research. Moreover, "information cocoons" have become the bottleneck of personalized marketing and recommendations. A number of studies report methods to accurately extract user interests, preferences, and value orientations from UGC and other behavior data; however, such methods are prone to generate repeated recommendations of large-scale similar information. In the future, how to overcome the problem of information cocoons and ensure the richness and accuracy of recommendation results should be considered by both scholars and businesses.

6.3. Other Emerging Problems

Although text mining can help to obtain an in-depth understanding and utilization of UGC in business and industry, some problems have emerged that warrant further attention.

Alongside business proliferation, business ethics is an indispensable subject. Some platforms use their collected big data to discriminate against consumers, so-called big data discriminatory pricing [133], seriously violating the legitimate rights and interests of consumers. Another phenomenon that largely impairs the rights of consumers is the leaking of private user data. Therefore, research focused on business ethics and exploring technologies and legislation to prevent the abuse of UGC data is urgently required.

In addition to text content, UGC exists in the form of images and videos, both on social sharing platforms, such as Instagram, TikTok, and Little Redbook, and on electronic and review websites, such as Dianping and Yelp. Therefore, elucidation of information extraction is important to obtain a comprehensive understanding of consumers. Nanne et al. [134] explored several applications of computer vision for brand marketing, as well as the usability of three pretrained, ready-to-use computer vision models, namely YOLOV2, Google Cloud Vision, and Clarifai, to automatically analyze visual brand-related UGC. The three models demonstrate the extent of usefulness of interpreting UGC images. With the development of computer version techniques, it will become possible to gain precise knowledge from images and videos and, which can be used to drive intelligent business practices.

7. Conclusions

The rapid development of e-commerce has resulted in UGC to playing a key role in business practices. Text mining of UGC can mitigate the purchase risk of latent buyers [55] and provide novel solutions for companies to obtain user feedback and conduct requirement analysis [81]. However, given the interdisciplinary nature of the domain, the coupling of linguistics, consumer behavior, economics, psychology, and computational science increases the difficulty of realizing specific business applications using text mining in an integrated manner [6]. In this study, we discussed the techniques of text mining and their applications in e-commerce from the perspective of solving practical problems and proposed prospective research directions in the field based on the interdisciplinary nature of the topic.

Our work makes four main contributions. (1) First, we summarized the general framework and methods of related research through a systematic review of the literature on text mining of UGC. (2) We introduced text mining techniques that are widely used, including semantic and sentiment analysis. Specifically, we analyzed the development status of semantic analysis in terms of text representation and semantic understanding. Then, the definition, development, and technical classification of sentiment analysis techniques were introduced in detail. (3) Furthermore, we discussed, in detail, the mainstream directions of text mining for business applications, ranging from high-quality UGC detection and consumer profiling to product enhancement and marketing. (4) Finally, we analyzed the research gaps with respect to these efforts and provided suggestions for future work. This study provides a comprehensive view of the field and can help to inspire valuable future research. We expect that research on text mining of UGC will continue to emerge to drive intelligent business.

Funding: This work was supported by the Chinese National Natural Science Foundation (No. 71871135 and No. 72271155).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ghani, N.A.; Hamid, S.; Hashem, I.A.T.; Ahmed, E. Social media big data analytics: A survey. *Comput. Hum. Behav.* **2019**, *101*, 417–428. [CrossRef]
2. Guo, Y.; Barnes, S.J.; Jia, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tour. Manag.* **2017**, *59*, 467–483. [CrossRef]
3. Soyeon, L.; Saerom, L. Does the dispersion of online review ratings affect review helpfulness? *Comput. Hum. Behav.* **2021**, *117*, 106670.
4. Gensler, S.; Völckner, F.; Egger, M.; Fischbach, K.; Schoder, D. Listen to Your Customers: Insights into Brand Image Using Online Consumer-Generated Product Reviews. *Int. J. Electron. Commer.* **2015**, *20*, 112–141. [CrossRef]
5. Qi, J.Y.; Zhang, Z.P.; Jeon, S.M.; Zhou, Y.Q. Mining customer requirements from online reviews: A product improvement perspective. *Inf. Manag.* **2016**, *53*, 951–963. [CrossRef]
6. Hou, T.J.; Yannou, B.; Leroy, Y.; Poirson, E. Mining customer product reviews for product development: A summarization process. *Expert Syst. Appl.* **2019**, *132*, 141–150. [CrossRef]

7. Rambocas, M.; Pacheco, B.G. Online sentiment analysis in marketing research: A review. *J. Res. Interact. Mark.* **2018**, *12*, 146–163. [CrossRef]
8. Stephen, A.T.; Toubia, O. Deriving Value from Social Commerce Networks. *J. Mark. Res.* **2010**, *47*, 215–228. [CrossRef]
9. Chiang, C.-T.; Yang, M.-H.; Koo, T.-L.; Liao, C.H. What Drives Customer Engagement Behavior? The Impact of User Participation from a Sociotechnical Perspective. *J. Electron. Commer. Res.* **2020**, *21*, 18.
10. Xia, H.S.; Yang, Y.T.; Pan, X.T.; Zhang, Z.P.; An, W.Y. Sentiment analysis for online reviews using conditional random fields and support vector machines. *Electron. Commer. Res.* **2020**, *20*, 343–360. [CrossRef]
11. Rangnekar, V.M.; Banker, D.D.; Jhala, H.I. Some phenotypic characteristics of group H plasmids from human isolates of Salmonella & Escherichia coli. *Indian J. Med. Res.* **1983**, *78*, 450–453. [PubMed]
12. Park, S.; Kim, H.M. Phrase Embedding and Clustering for Sub-Feature Extraction From Online Data. *J. Mech. Des.* **2022**, *144*, 10. [CrossRef]
13. Li, S.G.; Zhang, Y.Q.; Li, Y.M.; Yu, Z.X. The user preference identification for product improvement based on online comment patch. *Electron. Commer. Res.* **2021**, *21*, 423–444. [CrossRef]
14. Salloum, S.A.; Khan, R.; Shaalan, K. A Survey of Semantic Analysis Approaches. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), Settat, Morocco, 28–30 June 2020; pp. 61–70.
15. Meng, Y.; Yang, N.H.; Qian, Z.L.; Zhang, G.Y. What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *16*, 466–490. [CrossRef]
16. Burtch, G.; He, Q.; Hong, Y.; Lee, D. How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit. *Manag. Sci.* **2021**, *68*, 3175–3973. [CrossRef]
17. Li, Y.M.; Wei, B.G.; Liu, Y.H.; Yao, L.; Chen, H.; Yu, J.F.; Zhu, W.H. Incorporating knowledge into neural network for text representation. *Expert Syst. Appl.* **2018**, *96*, 103–114. [CrossRef]
18. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *arXiv* **2013**, arXiv:1310.4546v1.
19. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, Bejing, China, 22–24 June 2014; pp. 1188–1196.
20. Goldberg, Y. A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Intell. Res.* **2016**, *57*, 345–420. [CrossRef]
21. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805v2.
22. Ranaldi, L.; Fallucchi, F.; Zanzotto, F.M. Dis-Cover AI Minds to Preserve Human Knowledge. *Future Internet* **2022**, *14*, 10. [CrossRef]
23. Gupta, I.; Chatterjee, I.; Gupta, N. Latent Semantic Analysis based Real-world Application of Topic Modeling: A Review Study. In Proceedings of the 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 23–25 February 2022; pp. 1142–1149.
24. Egger, R.; Yu, J.N. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef] [PubMed]
25. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794v1.
26. Filieri, R.; Galati, F.; Raguseo, E. The impact of service attributes and category on eWOM helpfulness: An investigation of extremely negative and positive ratings using latent semantic analytics and regression analysis. *Comput. Hum. Behav.* **2021**, *114*, 106527. [CrossRef]
27. Cai, M.; Tan, Y.; Ge, B.; Dou, Y.; Huang, G.; Du, Y. PURA: A Product-and-User Oriented Approach for Requirement Analysis From Online Reviews. *IEEE Syst. J.* **2022**, *16*, 566–577. [CrossRef]
28. Araie, M.; Takase, M. Befunolol isomers and aqueous humor dynamic in man (author's transl). *Nippon Ganka Gakkai Zasshi* **1981**, *85*, 44–49.
29. Alrababah, S.A.A.; Gan, K.H.; Tan, T.P. Mining opinionated product features using WordNet lexicographer files. *J. Inf. Sci.* **2017**, *43*, 769–785. [CrossRef]
30. Yan, Z.J.; Xing, M.M.; Zhang, D.S.; Ma, B.Z. EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Inf. Manag.* **2015**, *52*, 850–858. [CrossRef]
31. Kang, Y.; Zhou, L.N. RubE: Rule-based methods for extracting product features from online consumer reviews. *Inf. Manag.* **2017**, *54*, 166–176. [CrossRef]
32. Chauhan, N.; Singh, P. Identifying the Opinion Orientation of Online Product Reviews at Feature Level: A Pruning Approach. *Int. J. Inf. Syst. Modeling Des.* **2017**, *8*, 92–111. [CrossRef]
33. Yong, S.; Asano, Y. Purpose-Feature Relationship Mining from Online Reviews towards Purpose-Oriented Recommendation. *IEICE Trans. Inf. Syst.* **2018**, *E101d*, 1021–1029. [CrossRef]
34. Kim, S.J.; Maslowska, E.; Malthouse, E.C. Understanding the effects of different review features on purchase probability. *Int. J. Advert.* **2018**, *37*, 29–53. [CrossRef]
35. Alamoudi, E.S.; Alghamdi, N.S. Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *J. Decis. Syst.* **2021**, *30*, 259–281. [CrossRef]

36. Wang, A.N.; Zhang, Q.; Zhao, S.Y.; Lu, X.N.; Peng, Z.L. A review-driven customer preference measurement model for product improvement: Sentiment-based importance-performance analysis. *Inf. Syst. E-Bus. Manag.* **2020**, *18*, 61–88. [CrossRef]
37. Zhang, H.; Rao, H.G.; Feng, J.Z. Product innovation based on online review data mining: A case study of Huawei phones. *Electron. Commer. Res.* **2018**, *18*, 3–22. [CrossRef]
38. Malik, M.S.I.; Hussain, A. Helpfulness of product reviews as a function of discrete positive and negative emotions. *Comput. Hum. Behav.* **2017**, *73*, 290–302. [CrossRef]
39. Cambria, E.; Schuller, B.; Xia, Y.Q.; Havasi, C. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [CrossRef]
40. Jin, J.; Ji, P.; Kwong, C.K. What makes consumers unsatisfied with your products: Review analysis at a fine-grained level. *Eng. Appl. Artif. Intell.* **2016**, *47*, 38–48. [CrossRef]
41. Wang, X.H.; Dong, S. Users' Sentiment Analysis of Shopping Websites Based on Online Reviews. *Appl. Math. Nonlinear Sci.* **2020**, *5*, 493–502. [CrossRef]
42. Asghar, M.Z.; Khan, A.; Zahra, S.R.; Ahmad, S.; Kundi, F.M. Aspect-based opinion mining framework using heuristic patterns. *Cluster Comput.* **2019**, *22*, S7181–S7199. [CrossRef]
43. Afzaal, M.; Usman, M.; Fong, A. Predictive aspect-based sentiment classification of online tourist reviews. *J. Inf. Sci.* **2019**, *45*, 341–363. [CrossRef]
44. Xu, H.P.; Zhang, Y.H.; Degroof, R. A Feature-Based Sentence Model for Evaluation of Similar Online Products. *J. Electron. Commer. Res.* **2018**, *19*, 320–335.
45. Wang, W.M.; Tian, Z.G.; Li, Z.; Wang, J.W.; Barenji, A.V.; Cheng, M.N. Supporting the construction of affective product taxonomies from online customer reviews: An affective-semantic approach. *J. Eng. Des.* **2019**, *30*, 445–476. [CrossRef]
46. Turner, P.S. Transformation of an unspecific chronic ulcer of the tongue into squamous cell carcinoma. *Quintessence Int. Dent. Dig.* **1983**, *14*, 703–707.
47. Liu, Y.; Bi, J.W.; Fan, Z.P. Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. *Inf. Fusion* **2017**, *36*, 149–161. [CrossRef]
48. Darwich, M.; Mohd Noah, S.A.; Omar, N.; Osman, N. Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. *J. Digit. Inf. Manag.* **2019**, *17*, 296. [CrossRef]
49. Peng, H.Y.; Cambria, E.; Hussain, A. A Review of Sentiment Analysis Research in Chinese Language. *Cogn. Comput.* **2017**, *9*, 423–435. [CrossRef]
50. Li, J.; Zhang, Y.; Li, J.; Du, J. The Role of Sentiment Tendency in Affecting Review Helpfulness for Durable Products: Nonlinearity and Complementarity. *Inf. Syst. Front.* **2022**, *158*, 1–19. [CrossRef]
51. Vinodhini, G.; Chandrasekaran, R.M. A sampling based sentiment mining approach for e-commerce applications. *Inf. Process. Manag.* **2017**, *53*, 223–236. [CrossRef]
52. Su, Y.; Shen, Y. A Deep Learning-Based Sentime.ent Classification Model for Real Online Consumption. *Front. Psychol.* **2022**, *13*, 886982. [CrossRef]
53. Sun, Q.; Niu, J.W.; Yao, Z.; Yan, H. Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level. *Eng. Appl. Artif. Intell.* **2019**, *81*, 68–78. [CrossRef]
54. Munikar, M.; Shakya, S.; Shrestha, A. Fine-grained Sentiment Classification using BERT. In Proceedings of the 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 5 November 2019; Volume 1, pp. 1–5.
55. Saumya, S.; Singh, J.P.; Baabdullah, A.M.; Rana, N.P.; Dwivedi, Y.K. Ranking online consumer reviews. *Electron. Commer. Res. Appl.* **2018**, *29*, 78–89. [CrossRef]
56. Banerjee, S.; Bhattacharyya, S.; Bose, I. Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. *Decis. Support. Syst.* **2017**, *96*, 17–26. [CrossRef]
57. Hu, N.; Bose, I.; Koh, N.S.; Liu, L. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decis. Support Syst.* **2012**, *52*, 674–684. [CrossRef]
58. Filieri, R.; Hofacker, C.F.; Alguezaui, S. What makes information in online consumer reviews diagnostic over time? The role of review relevancy, factuality, currency, source credibility and ranking score. *Comput. Hum. Behav.* **2018**, *80*, 122–131. [CrossRef]
59. Li, M.X.; Huang, L.Q.; Tan, C.H.; Wei, K.K. Helpfulness of Online Product Reviews as Seen by Consumers: Source and Content Features. *Int. J. Electron. Commer.* **2013**, *17*, 101–136. [CrossRef]
60. Choi, H.S.; Leon, S. An empirical investigation of online review helpfulness: A big data perspective. *Decis. Support. Syst.* **2020**, *139*, 113403. [CrossRef]
61. Yagci, I.A.; Das, S. Measuring design-level information quality in online reviews. *Electron. Commer. Res. Appl.* **2018**, *30*, 102–110. [CrossRef]
62. Chatterjee, S. Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach. *Int. J. Hosp. Manag.* **2020**, *85*, 9. [CrossRef]
63. Zhang, D.; Zhou, L.; Kehoe, J.L.; Kilic, I.Y. What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews. *J. Manag. Inform. Syst.* **2016**, *33*, 456–481. [CrossRef]
64. Wu, C.J.; Mai, F.; Li, X.L. The effect of content depth and deviation on online review helpfulness: Evidence from double-hurdle model. *Inf. Manag.* **2021**, *58*, 103408. [CrossRef]

65. Li, J.; Zhan, L.J. Online Persuasion: How the Written Word Drives WOM Evidence from Consumer-Generated Product Reviews. *J. Advert. Res.* **2011**, *51*, 239–257. [CrossRef]
66. Kang, Y.; Zhou, L.N. Helpfulness Assessment of Online Reviews: The Role of Semantic Hierarchy of Product Features. *ACM Trans. Manag. Inf. Syst.* **2019**, *10*, 18. [CrossRef]
67. Nilsson, M.O.; Hultin, T. Analysis of the membrane-associated poly(A)+RNA in the cytoplasm of dormant Artemia cysts by DNA excess hybridization. Evidence for a nuclear origin. *Biochim. Biophys. Acta* **1982**, *696*, 253–259. [CrossRef]
68. Lee, J.H.; Jung, S.H.; Park, J. The role of entropy of review text sentiments on online WOM and movie box office sales. *Electron. Commer. Res. Appl.* **2017**, *22*, 42–52. [CrossRef]
69. Cezar, A.; Ogut, H. Analyzing conversion rates in online hotel booking The role of customer reviews, recommendations and rank order in search listings. *Int. J. Contemp. Hosp. Manag.* **2016**, *28*, 286–304. [CrossRef]
70. Suryadi, D.; Kim, H. A Systematic Methodology Based on Word Embedding for Identifying the Relation Between Online Customer Reviews and Sales Rank. *J. Mech. Des.* **2018**, *140*, 12. [CrossRef]
71. Dash, A.; Zhang, D.S.; Zhou, L.N. Personalized Ranking of Online Reviews Based on Consumer Preferences in Product Features. *Int. J. Electron. Commer.* **2021**, *25*, 29–50. [CrossRef]
72. Paul, H.; Nikolaev, A. Fake review detection on online E-commerce platforms: A systematic literature review. *Data Min. Knowl. Discov.* **2021**, *35*, 1830–1881. [CrossRef]
73. Kauffmann, E.; Peral, J.; Gil, D.; Ferrandez, A.; Sellers, R.; Mora, H. A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Ind. Mark. Manag.* **2020**, *90*, 523–537. [CrossRef]
74. Saumya, S.; Singh, J.P. Spam review detection using LSTM autoencoder: An unsupervised approach. *Electron. Commer. Res.* **2022**, *22*, 113–133. [CrossRef]
75. Li, S.G.; Liu, F.; Lu, H.Y.; Zhang, Y.Q.; Li, Y.M.; Yu, Z.X. Product family lean improvement based on matching deep mining of customer group preference. *Res. Eng. Des.* **2021**, *32*, 469–488. [CrossRef]
76. Qiu, F.; Cho, J. Automatic identification of user interest for personalized search. In Proceedings of the 15th International Conference on World Wide Web, Edinburgh, UK, 23–26 May 2006; pp. 727–736.
77. Hong, W.; Zheng, C.Y.; Wu, L.H.; Pu, X.J. Analyzing the Relationship between Consumer Satisfaction and Fresh E-Commerce Logistics Service Using Text Mining Techniques. *Sustainability* **2019**, *11*, 3570. [CrossRef]
78. Xu, X. What are customers commenting on, and how is their satisfaction affected? Examining online reviews in the on-demand food service context. *Decis. Support Syst.* **2021**, *142*, 113467. [CrossRef]
79. Bi, J.W.; Liu, Y.; Fan, Z.P.; Zhang, J. Exploring asymmetric effects of attribute performance on customer satisfaction in the hotel industry. *Tour. Manag.* **2020**, *77*, 104006. [CrossRef]
80. Liu, H.Y.; He, J.; Wang, T.T.; Song, W.T.; Du, X.Y. Combining user preferences and user opinions for accurate recommendation. *Electron. Commer. Res. Appl.* **2013**, *12*, 14–23. [CrossRef]
81. Chen, R.Y.; Wang, Q.L.; Xu, W. Mining user requirements to facilitate mobile app quality upgrades with big data. *Electron. Commer. Res. Appl.* **2019**, *38*, 11. [CrossRef]
82. Ekutsu, M.; Floras, P.; Macrez, P.; Bidabe, A.M.; Caille, J.M. Sedation for tomodensitometric examination in children. *Cah. Anesthesiol.* **1984**, *32*, 375–378.
83. Kasper, G.; de Siqueira Braga, D.; Martins, D.M.L.; Hellingrath, B. User profile acquisition: A comprehensive framework to support personal information agents. In Proceedings of the 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Arequipa, Peru, 8–10 November 2017; pp. 1–6.
84. Chen, H.C.; Chiang, R.H.L.; Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Q.* **2012**, *36*, 1165–1188. [CrossRef]
85. Sun, H.; Guo, W.; Shao, H.Y.; Rong, B. Dynamical mining of ever-changing user requirements: A product design and improvement perspective. *Adv. Eng. Inform.* **2020**, *46*, 101174. [CrossRef]
86. Zhou, S.H.; Qiao, Z.L.; Du, Q.Z.; Wang, G.A.; Fan, W.G.; Yan, X.B. Measuring Customer Agility from Online Reviews Using Big Data Text Analytics. *J. Manag. Inform. Syst.* **2018**, *35*, 510–539. [CrossRef]
87. Kim, C.; Na, Y. Consumer reviews analysis on cycling pants in online shopping malls using text mining. *Fash. Text.* **2021**, *8*, 38. [CrossRef]
88. Bi, J.W.; Liu, Y.; Fan, Z.P.; Cambria, E. Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. *Int. J. Prod. Res.* **2019**, *57*, 7068–7088. [CrossRef]
89. Hou, T.J.; Yannou, B.; Leroy, Y.; Poirson, E. Mining Changes in User Expectation Over Time From Online Reviews. *J. Mech. Des.* **2019**, *141*, 10. [CrossRef]
90. Francesco, G.; Roberta, G. Cross-country analysis of perception and emphasis of hotel attributes. *Tour. Manag.* **2019**, *74*, 24–42. [CrossRef]
91. Chatterjee, S.; Goyal, D.; Prakash, A.; Sharma, J. Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *J. Bus. Res.* **2021**, *131*, 815–825. [CrossRef]
92. Xia Liu, A.; Yilin, L.; Sean Xin, X. Assessing the Unacquainted: Inferred Reviewer Personality and Review Helpfulness. *MIS Q.* **2021**, *45*, 1113–1148.

93. Donthu, N.; Kumar, S.; Pandey, N.; Pandey, N.; Mishra, A. Mapping the electronic word-of-mouth (eWOM) research: A systematic review and bibliometric analysis. *J. Bus. Res.* **2021**, *135*, 758–773. [CrossRef]
94. Deng, L.; Gao, J.; Vuppalapati, C. Building a Big Data Analytics Service Framework for Mobile Advertising and Marketing. In Proceedings of the 2015 IEEE First International Conference on Big Data Computing Service and Applications, Redwood City, CA, USA, 30 March–2 April 2015; pp. 256–266.
95. Baum, D.; Spann, M. The Interplay Between Online Consumer Reviews and Recommender Systems: An Experimental Analysis. *Int. J. Electron. Commer.* **2014**, *19*, 129–161. [CrossRef]
96. Hernandez-Rubio, M.; Cantador, I.; Bellogin, A. A comparative analysis of recommender systems based on item aspect opinions extracted from user reviews. *User Model. User-Adapt. Interact.* **2019**, *29*, 381–441. [CrossRef]
97. Forhad, M.S.A.; Arefin, M.S.; Kayes, A.S.M.; Ahmed, K.; Chowdhury, M.J.M.; Kumara, I. An Effective Hotel Recommendation System through Processing Heterogeneous Data. *Electronics* **2021**, *10*, 21. [CrossRef]
98. Wang, W.; Wang, H.W. Opinion-enhanced collaborative filtering for recommender systems through sentiment analysis. *New Rev. Hypermedia Multimed.* **2015**, *21*, 278–300. [CrossRef]
99. Yuqiao, Y.; Weiping, Z.; Fangdan, L.; Hongli, Z. Personalized information recommendation simulation system based on compound recommendation algorithm—A research tool to study the push effect of algorithm. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *740*, 012167. [CrossRef]
100. DeLong, K.A.; Troyer, M.; Kutas, M. Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Lang. Linguist. Compass* **2014**, *8*, 631–645. [CrossRef] [PubMed]
101. Oliveira, A.S.; Renda, A.I.; Correia, M.B.; Antonio, N. Hotel customer segmentation and sentiment analysis through online reviews: An analysis of selected European markets. *Tour. Manag. Stud.* **2022**, *18*, 29–40. [CrossRef]
102. Yuan, H.; Xu, W.; Li, Q.; Lau, R. Topic sentiment mining for sales performance prediction in e-commerce. *Ann. Oper. Res.* **2018**, *270*, 553–576. [CrossRef]
103. Skoric, M.; Poor, N.; Achananuparp, P.; Lim, E.-P.; Jiang, J. Tweets and votes: A study of the 2011 singapore general election. In Proceedings of the 2012 45th Hawaii International Conference On System Sciences, Maui, HI, USA, 4–7 January 2012; pp. 2583–2591.
104. Zou, H.; Tang, X.; Xie, B.; Liu, B. Sentiment classification using machine learning techniques with syntax features. In Proceedings of the 2015 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 7–9 December 2015; pp. 175–179.
105. Ghose, A.; Ipeirotis, P.G.; Li, B.B. Modeling Consumer Footprints on Search Engines: An Interplay with Social Media. *Manag. Sci.* **2019**, *65*, 1363–1385. [CrossRef]
106. Xu, Y.B.; Yang, Y.J.; Han, J.Y.; Wang, E.; Ming, J.C.; Xiong, H. Slanderous user detection with modified recurrent neural networks in recommender system. *Inf. Sci.* **2019**, *505*, 265–281. [CrossRef]
107. Zheng, X.L.; Zhu, S.; Lin, Z.X. Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach. *Decis. Support. Syst.* **2013**, *56*, 211–222. [CrossRef]
108. Li, J.P.; Yao, Y.H.; Xu, Y.J.; Li, J.Y.; Wei, L.; Zhu, X.Q. Consumer's risk perception on the Belt and Road countries: Evidence from the cross-border e-commerce. *Electron. Commer. Res.* **2019**, *19*, 823–840. [CrossRef]
109. Chen, F.; Xia, J.X.; Gao, H.H.; Xu, H.H.; Wei, W. TRG-DAtt: The Target Relational Graph and Double Attention Network Based Sentiment Analysis and Prediction for Supporting Decision Making. *ACM Trans. Manag. Inf. Syst.* **2022**, *13*, 1–25. [CrossRef]
110. Wang, H.W.; Gao, S.; Yin, P.; Liu, J.N.K. Competitiveness analysis through comparative relation mining Evidence from restaurants' online reviews. *Ind. Manag. Data Syst.* **2017**, *117*, 672–687. [CrossRef]
111. Song, C.; Zheng, L.; Shan, X. An analysis of public opinions regarding Internet-famous food: A 2016–2019 case study on Dianping. *Br. Food J.* 2022, *ahead-of-print*. [CrossRef]
112. Neirotti, P.; Raguseo, E.; Paolucci, E. Are customers' reviews creating value in the hospitality industry? Exploring the moderating effects of market positioning. *Int. J. Inf. Manag.* **2016**, *36*, 1133–1143. [CrossRef]
113. Kumar, S.; Chowdary, C.R. Semantic model to extract tips from hotel reviews. *Electron. Commer. Res.* **2020**, *259*, 1–19. [CrossRef]
114. Zhao, Y.B.; Xu, X.; Wang, M.S. Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *Int. J. Hosp. Manag.* **2019**, *76*, 111–121. [CrossRef]
115. Raguseo, E.; Neirotti, P.; Paolucci, E. How small hotels can drive value their way in infomediation. The case of 'Italian hotels vs. OTAs and TripAdvisor'. *Inf. Manag.* **2017**, *54*, 745–756. [CrossRef]
116. Liu, M.T.C.; Liu, Y.D.; Mo, Z.Y.; Ng, K.L. Using text mining to track changes in travel destination image: The case of Macau. *Asia Pac. J. Mark. Logist.* **2021**, *33*, 373–395. [CrossRef]
117. Zhang, T.X.; He, Z.; Zhao, X.J.; Qu, L. Joint monit.toring of post-sales online review processes based on a distribution-free EWMA scheme. *Comput. Ind. Eng.* **2021**, *158*, 107372. [CrossRef]
118. Mohawesh, R.; Tran, S.; Ollington, R.; Xu, S.X. Analysis of concept drift in fake reviews detection. *Expert Syst. Appl.* **2021**, *169*, 114318. [CrossRef]
119. Bilal, M.; Marjani, M.; Hashem, I.A.T.; Malik, N.; Lali, M.I.U.; Gani, A. Profiling reviewers' social network strength and predicting the "Helpfulness" of online customer reviews. *Electron. Commer. Res. Appl.* **2021**, *45*, 101026. [CrossRef]
120. Nakayama, M.; Wan, Y. The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews. *Inf. Manag.* **2019**, *56*, 271–279. [CrossRef]

121. Sandulescu, V.; Ester, M. Detecting Singleton Review Spammers Using Semantic Similarity. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 971–976.
122. Rui, H.; Liu, Y.; Whinston, A. Whose and what chatter matters? The effect of tweets on movie sales. *Decis. Support Syst.* **2013**, *55*, 863–870. [CrossRef]
123. Jagiripu, I.P.; Mishra, P.K.; Saini, A.; Biswal, A. Testing the impact of uncertainty reducing reviews in the prediction of cross domain social media pages ratings. *J. Indian Bus. Res.* **2022**, *14*, 150–166. [CrossRef]
124. Ma, Y.; Chen, G.Q.; Wei, Q. Finding users preferences from large-scale online reviews for personalized recommendation. *Electron. Commer. Res.* **2017**, *17*, 3–29. [CrossRef]
125. Yang, M.; Ma, Y.; Nie, J. Research on a Personalized Recommendation Algorithm. *Int. J. Grid Distrib. Comput.* **2017**, *10*, 123–136. [CrossRef]
126. Wang, J.Q.; Zhang, X.; Zhang, H.Y. Hotel recommendation approach based on the online consumer reviews using interval neutrosophic linguistic numbers. *J. Intell. Fuzzy Syst.* **2018**, *34*, 381–394. [CrossRef]
127. Huang, Z.; Benyoucef, M. From e-commerce to social commerce: A close look at design features. *Electron. Commer. Res. Appl.* **2013**, *12*, 246–259. [CrossRef]
128. Salloum, S.A.; Al-Emran, M.; Monem, A.A.; Shaalan, K. A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives. *Adv. Sci. Technol. Eng. Syst. J.* **2017**, *2*, 127–133.
129. Hotho, A.; Nurnberger, A.; Paaß, G.; Augustin, S. A Brief Survey of Text Mining. *ResearchGate* **2005**, *37*, 19–62.
130. Hsiao, Y.H.; Chen, M.C.; Liao, W.C. Logistics service design for cross-border E-commerce using Kansei engineering with text-mining-based online content analysis. *Telemat. Inform.* **2017**, *34*, 284–302. [CrossRef]
131. Zhang, C.; Tian, Y.X.; Fan, L.W.; Li, Y.H. Customized ranking for products through online reviews: A method incorporating prospect theory with an improved VIKOR. *Appl. Intell.* **2020**, *50*, 1725–1744. [CrossRef]
132. Li, M. Research on Extraction of Useful Tourism Online Reviews Based on Multimodal Feature Fusion. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2021**, *20*, 16. [CrossRef]
133. Liu, W.; Long, S.; Xie, D.; Liang, Y.; Wang, J. How to govern the big data discriminatory pricing behavior in the platform service supply chain?An examination with a three-party evolutionary game model. *Int. J. Prod. Econ.* **2021**, *231*, 107910. [CrossRef]
134. Nanne, A.J.; Antheunis, M.L.; van der Lee, C.G.; Postma, E.O.; Wubben, S.; van Noort, G. The Use of Computer Vision to Analyze Brand-Related User Generated Image Content. *J. Interact. Mark.* **2020**, *50*, 156–167. [CrossRef]