ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

Advances in Decision Sciences

Volume 30
Issue 2
June 2026

亞洲大學
ASIA UNIVERSITY
SCIENTIFIC & BUSINESS WORLD

Published by Asia University, Taiwan and Scientific and Business World

Review Helpfulness to Support Business: Identifying Fake Reviews from User-Generated Content Using Random Forest

Syed Imran Abbas Qazmi
Lincoln University College, Malaysia
***Corresponding author* Email:** siaqazmi@lincoln.edu.my

Midhun Chakkaravarthy
Faculty of Computer Science and Multimedia,
Lincoln University College, Malaysia
Email: midhun@lincoln.edu.my

Syed Hassan Raza
School of Media and Communication, Taylor's University,
47500 Subang Jaya, Selangor, Malaysia
Email: Hassanraza.syed@taylors.edu.my

Farrah Aslam
Department of Information Sciences, University of Education Lahore,
Jauharabad Campus, Pakistan
Email: farrah.aslam@ue.edu.pk

Shahbaz Aslam
Department of Media and Communication Studies,
Comsat University, Lahore 54000, Pakistan.
Email:shahbaz_vu@yahoo.com

Moneeba Iftikhar
Department of Mass Communication,
Lahore College for Women University, Lahore 54000, Pakistan.
Email: moneeba.iftikhar@lcwu.edu.pk

Received: January 18, 2025; First Revision: June 4, 2025;

Last Revision: March 15, 2026; Accepted: March 25, 2026;

Published: March 28, 2026

Abstract

Purpose: Valid and Helpful reviews on an e-commerce platform provide important information regarding customers' perception of a product, which is crucial to the existence and growth of any business. False reviews, which are created to tarnish a product's image through spam fraudulently, continue to be a significant challenge for all e-commerce platforms. Another challenge remains in identifying helpful review content on the platform that can significantly alter a customer's opinion of a product. Hence, the increasing prevalence of fake and unhelpful reviews compromises the credibility of online reviews, resulting in information overload and a misleading consumer decision-making process. Motivated by this challenge, this study aims to develop an automated system capable of retaining only applicable and valid reviews to support the identification of customer needs, which is a valuable area of research.

Design/methodology/approach: This study involves three main aspects: helpfulness classification, fake review detection, and topic identification on various categories of the Amazon Dataset. The model leveraged a feature set that included the sentiment polarity of the review in detail, word count indicating the length of feedback, word diversity in the review, comprehension analysis of parts of speech in the review reflecting its grammatical structure and complexity, and authenticity metrics. Moreover, for helpful review classification, the utilized features included review and product metadata, review content informativeness score encoded with the help of Sentence Bidirectional Encoder Representations from Transformers (SBERT), and reviewer attributes. A topic extraction model has been implemented that leverages Gemini to extract sentiment-based topic analysis over reviews.

Findings: The study provides useful reviews classification over 6 different Amazon categories using a Random Forest classifier (RFC) by achieving 94% accuracy, precision, and F1-Score, a recall of 93%, and an AUC Score of 98%. While the Gradient Boosting classifier yielded comparable performance with an AUC Score of 98% and 94% accuracy, precision, recall, and F1-Score. For fake reviews detection in the Toys and Games category, the RFC achieved 85% accuracy, 86% precision, a 97% recall, 91% F1-Score, and 79% AUC Score. The findings indicate that combining textual, semantic, reviewer, and product-level features can improve the reliability of review quality assessment. Finally, to enhance the decision-making process for businesses, a topic extraction model utilizing the Gemini tool has been employed to extract significant topics from valid and helpful reviews, categorizing them separately into negative and positive reviews, thereby gaining nuanced insights into customer feedback.

Originality/value: Unlike prior studies that either examine review helpfulness or fake review detection in isolation, this study moves beyond single-task and small-sample-based approaches. Our proposed framework offers a comprehensive analysis of patterns in reviews across e-commerce platforms, thereby enhancing brands' ability to integrate customer needs and expectations into future marketing communications and advertising campaigns. This study contributes to Decision Sciences by proposing a data-driven two-stage framework that retains only helpful and valid reviews to enhance content quality, thereby practically supporting better decision-making by content moderation, reducing information overload, and improving consumer trust in reviews.

Keywords: Text Similarity, BERT Embeddings, Fake Reviews, Machine Learning, Review Helpfulness, Random Forest Classifier.

JEL Classifications: L81, O33, L86, C38, C45

1. Introduction

The early decades of the 21st century have witnessed the rise of Electronic Commerce (e-commerce). With the increase in digital business activities, including e-commerce platforms, online customer-generated content is critical for businesses to develop effective marketing communication campaigns (Nguyen et al., 2025). In modern e-commerce environments, online reviews serve as the most influential source of information, guiding both consumer and managerial strategies (Kübler et al., 2024; Ranfagni & Rosati, 2023). The feedback and reviews about products or services range from pricing, product features, and the transparency of information available in reviews(Kübler et al., 2024; Ranfagni & Rosati, 2023). Online feedback content is being rapidly generated at a pace that overwhelms consumers and businesses (U. Singh et al., 2022).

Despite having much assistance, customers also have a strong reservation about selecting the right product because, in online shopping, they cannot see or touch the product to assess its quality (Ranfagni & Rosati, 2023). Therefore, customers' online reviews are the primary source for evaluating product quality. Similarly, marketers struggle to derive actionable insights from a large volume of received feedback on their products, which can delay decision-making. Therefore, reviews play a dual role in marketers' communication strategies: first, they provide insight into customers' needs and expectations, and second, they help them clarify their stance and communicate effectively in future marketing communications(Kübler et al., 2024).

It becomes difficult to read many reviews to understand, analyze, and identify customers' needs (Kübler et al., 2024). On the other hand, online reviews motivate customers to keep reading the feedback to get advice and information on in-fashion online products (Qiu & Zhang, 2024). However, it reduces the positive opinion of customers when the genuine product differs from the one being presented online (H. Singh et al., 2023). Past research has revealed several interesting facets, but these findings also have some limitations. For instance, earlier research has focused on online reviews in specific fields, such as the fashion sector (Camacho-Otero et al., 2019). The findings may have limitations due to the limited scope of the framework used to categorize online reviews, which restricts their implications for business and marketing communication planning.

To this end, scholars have only assessed the determining factors of online review helpfulness, such as review intensity, readability, and rating (Deldjoo et al., 2023). Recent research indicates that length (Racherla & Friske, 2012), readability, and sentiment (Yin et al., 2016) are key components of the helpfulness of reviews. Hence, scarce research has been conducted on other dimensions of helpful reviews on online platforms. To our knowledge, product metadata incorporated into understanding the usefulness of reviews has been overlooked in the digital business literature. This research is motivated to address this void and explore the utilization of review metadata, reviewer credibility features, and review informativeness in relation to product metadata.

Cosine resemblance is generally adopted in the script cataloging domain due to its computational effectiveness and reliable presentation (Park et al., 2020; Singh & Garg, 2024). It can be used to verify whether some reviews contain high similarity rates with the product information, which may indicate the review to be more revealing and enlightening. While comparing different models used in text classification and sentiment analysis, it was found that domain-specific word embedding models enhance text analytics performance (Asudani et al., 2023). To classify the reviews, we further explore similarity measures that match the predefined standards after performing vectorization on the reviews (Shahmirzadi et al., 2019).

Previous research has also established that customer feedback is pivotal in business development because it provides valuable insights that drive product improvement, enhance customer satisfaction, and inform strategic decision-making (Lin, 2020). With the help of online reviews, the company enhances its products based on customer preferences, experiences, and customers' purchase decisions (Ren & Hong, 2019). When a review frequently occurs, it may be an indicator of a specific business demand that needs to be addressed and predicts the customers' needs. The analysis of these reviews proves to be very helpful for business development. However, it is not easy to recognize the helpful and valuable reviews in the vast collection.

Furthermore, false reviews present online may artificially modify the view of merchandise. An investigation by Canvas8 in 2020 highlighted the impact of reviews on consumers' decision-making, finding that 9 out of 10 consumers decide to buy a product, emphasizing the need for authenticity in feedback (Daniels, 2022). Bogus reviews can misinform consumers and business leaders, impacting their decision-making (INC42, 2024). Detecting fake reviews is, therefore, critical for academics studying online trust of reviews and for practitioners aiming to maintain the credibility of their platforms. Motivated by this challenge, this study proposes the extraction of linguistic and lexical features from review texts to determine the credibility of online reviews.

This research presents a helpful review classification as a pre-processing technique based on product data, reviewer data, review text similarity, and textual content analysis techniques. Moreover, the false review detection classification is prepared using Natural Language Processing. This research has applied different machine learning algorithms, Natural language processing, and advanced script analysis to distinguish helpful and unhelpful reviews and monitor false reviews. Natural Language Processing (NLP) has influenced language-constructed methods to investigate texts, comprehend their meaning, and extract information (Hannan et al., 2012). A limited study has explored both fake review and usefulness classification as a pre-processing step in decision-making. This study advances the field of decision sciences by investigating the role of resemblance in reviews of manufactured goods and merchandise descriptive content, along with reviewer history, in making a review helpful to consumers. Business and consumer decision-making is supported by beneficial information from reliable customer experiences.

Relevant and dependable constructive reviews provide information about the features of products that are well-received among consumers. This enables businesses to quickly assess overall consumer satisfaction and identify areas that require immediate attention. Recognizing the significance of criticism in decision-

making and identifying false reviews is crucial for promoting fair consumer practices and supporting sustainable e-commerce (Duma et al., 2024).

The study proposes a two-stage framework that identifies reviews customers find helpful and filters out false ones, thereby aiding in e-commerce decision-making and helping to comprehend buyer desires to improve and increase their merchandise and facilities. Academics can leverage the research findings to explore further on consumer behavior, review credibility, and improve trust in online platforms. Ultimately, this approach enables businesses to practically maintain a competitive edge and foster customer loyalty by continually improving their products and services based on valuable customer insights. We propose a solution to the challenge of interpreting every review and weighing the characteristics of a commercial product through feedback, which is consequential in the problem of information excess (U. Singh et al., 2022). The central contributions of the study are:

- To develop an automatic helpfulness assessment structure using customer review content, reviewer, and product attributes to identify helpful content.
- To filter out fake reviews from the review pools for further analysis, to support the customer needs identification system with only valid content.
- Implementation of topic identification over product user reviews, leveraging advanced AI tools like Gemini
- To leverage sentiment analysis and topic analysis to identify strengths by major topics in reviews having positive sentiments and weaknesses from major discussed topics among negative reviews.

The organization of the remainder of this paper is as follows: Section 2 presents the literature review, Section 3 presents the theory and hypothesis development section. Section 4 describes our proposed methodology, which begins by outlining our data collection approach, then details the experimental setup, the selected features, and the selection of the machine learning algorithm. Furthermore, the results are presented in Section 5, along with a discussion that provides an analysis of the results and their practical implications. Lastly, Section 6 provides the conclusion and discusses future directions.

2. Literature Review

Identifying customer needs is essential for a business to succeed in the market. People tend to trust reviews with a higher percentage of helpful votes, believing them to be more reliable than those without helpful votes (X. Sun et al., 2019). Many studies utilize machine learning to automatically analyze customer needs from social media data. Some of the research work is discussed in this section.

It has been noted that over 90% of prospective customers globally trust customer reviews and experiences (Zheng, 2021). Past business research has reported several noteworthy findings regarding online reviews, including impact of customer review images (Kübler et al., 2024), company brand success (Ranfagni & Rosati, 2023), customer trust in product quality (Sung et al., 2023), product and country image (Nguyen et al., 2025), and purchasing decisions (Qiu & Zhang, 2024). In this scenario, online reviews justify the

explicit scholarly attention, as they consolidate previously documented findings and highlight their role in diverse domains of business research, while also providing critical information to guide digital business practitioners.

Recent research focusing on the roles of product certainty and review helpfulness examines how online reviews and product information influence customer satisfaction. Previous studies offer valuable insights for e-commerce corporations and firms to align their business models with consumer expectations and enhance customer satisfaction (Changchit & Klaus, 2020). The findings reveal that review relevance affects the impact of online product reviews almost as strongly as review trustworthiness. Additionally, studies show mixed affiliations between review credibility and its drivers, with a positive link to trustworthiness but an undesirable one to expertise (Mumuni et al., 2020). The growing variety, volume, and speed specifically for fashion manufacturing pose a substantial challenge in the fashion industry, making it problematic for customers to choose which merchandise to purchase. In addition, fashion is an intrinsically subjective cultural belief and an ensemble of clothing pieces that keep a coherent style (Shirkhani et al., 2023). Literature suggests that appraisal usefulness can be well defined as the customer's awareness of the appraisal's ability to aid them in forming informed purchase decisions (Chatterjee, 2025). Table 1 presents a summary of the recent scholarly contributions that investigate various dimensions of online reviews.

Table 1. Recent Literature Matrix

Literature	Focus	Method	Background/Industry
(Nguyen et al., 2025)	Online Reviews and Country product image and purchase intention	Quantitative	Foreign Products/ E-commerce
(Kübler et al., 2024)	The effect of review images on review helpfulness	Experiment	Online retailing/ E-commerce
(Ranfagni & Rosati, 2023)	Online brand reputation	Interdisciplinary approach	Hospitality
(Sung et al., 2023)	Consumer Trust based on Online reviews	Empirical study	E-commerce
(Qiu & Zhang, 2024)	Cultural Context of Online Reviews Influences Purchase Intention	Meta-Analysis	E-commerce
(Deldjoo et al., 2023)	AI-driven Fashion recommender	Systematic review	Fashion (E-commerce)
(H. Singh et al., 2023)	Gratifications and credibility of reviews as drivers of buying	Survey	Fashion (E-commerce)

Note: The compilation supports the theoretical and empirical foundation of our research by synthesizing key themes and application contexts. It highlights the value of insights derived from user-generated content in enhancing our understanding of review helpfulness and credibility.

Therefore, feature engineering is essential for improving model performance. Features such as review ratings, verified purchase status, and text similarity between review text and product details are used to enhance model prediction accuracy. These features help to understand the review's usefulness and level of customer satisfaction. The research on reviews primarily relies on review text (Fan et al., 2018), earlier studies employed methods such as cosine similarity and Euclidean distance to measure text similarity.

Cosine similarity measures the closeness of two vectors by calculating the angle between them. A value of 1 represents the perfect alignment, 0 represents orthogonality, and -1 means opposite (Jyoti & Singh, 2015).

BERT learns deep bidirectional, contextualized text representations from large-scale unlabeled corpora, effectively overcoming the limitations of previous static word embedding techniques (Devlin et al., 2019; Vaswani et al., 2017). By considering bidirectional context, it is capable of capturing semantic relationships useful for text similarity and classification tasks (C. Sun et al., 2019; Yoo & Jeong, 2020). A review of BERT's operation and applications, comparing it with similar models, proved its superiority for text analysis (Koroteev, 2021). Another study fine-tuned BERT for categorizing helpful and unhelpful reviews, comparing its performance to traditional bag-of-words methods. It also examines the impact of varying sequence lengths on BERT's efficiency in predicting review helpfulness using Yelp Open Dataset reviews. Studies have reported high precision and recall rates using BERT to identify helpful reviews, empowering commerce to adapt their products and services more effectively to customer demands (Bilal & Almazroi, 2023). Combining BERT embeddings and machine learning techniques offers a robust framework for identifying helpful reviews. Various solutions for evaluating helpfulness using machine learning algorithms, such as K Nearest Neighbor (KNN) and Support Vector Machine (SVM) were studied.

Reviews of product features, such as size and storage, are typically similar and do not provide helpful information. Also, a review needs a sufficient number of votes to be considered helpful by customers. Four regression models were applied to determine the connection between review helpfulness and informativeness, which was then used to create a threshold to decide review helpfulness. They also examined how the threshold could differ for search and experience products (X. Sun et al., 2019).

Another system used sentiment analysis to predict the helpfulness of a review, examining the positivity, negativity, and emotions in reviews (Lin, 2020). Various studies have considered deep learning approaches, such as embedding gated Convolutional Neural Networks (CNNs) (Chen et al., 2019) and multitask neural learning (Fan et al., 2018), which extract text features from reviews to understand their meaning. To determine helpfulness, these models relied solely on review text and ignored product details. Additionally, some methods have used lexical, structural, and semantic analysis, word embeddings, and Flesch reading ease to predict review helpfulness using decision trees as a classifier. Their work focused on only three categories from the Amazon dataset (Enamul Haque et al., 2018).

The spread of fake reviews poses a threat to the credibility of online platforms (Periasamy et al., 2024). Manual detection of fake reviews is challenging when generated by machines. Comparing a baseline SVM algorithm with an OpenAI detection model revealed that fine-tuned AI models outperform the baseline. Another model was developed using fine-tuning of RoBERTa for a text classification task. The model achieved a 97% F1 score, precision, and recall using an 80/20 split for training and testing (Salminen et al., 2022).

Abd and Hussein discussed numerous machine-learning approaches for detecting fake reviews, including supervised, semi-supervised, and unsupervised learning. Another study employed the K Nearest Neighbor

algorithm to classify reviews based on features such as understanding, age, and number of posts, posting time, location, product type, and feedback. Reviews were obtained from the available datasets, which included many features. For instance, understanding age, number of posts, the post's time, place, product-based, and purchaser feedback (Paul & Nikolaev, 2021).

Most previous studies solely focused on reviewing text or metadata to support helpfulness prediction. We propose review text, authors, and product metadata-based features to improve the results. To detect fake reviews, our proposed work suggests a lexical and linguistic analysis of only review text without adding the complexity of the author's history. Here, we implement two different algorithms that work sequentially to filter out unhelpful and fake reviews.

3. Theory and Hypothesis Development

Prior literature suggests that higher helpfulness votes make a review seem trustworthy (X. Sun et al., 2019). Additionally, the review informativeness, sentiment polarity, and helpful votes are collectively related to the quality of a review (Chatterjee, 2025; Lin, 2020; Mumuni et al., 2020). This study extends existing research by analyzing linguistic features for predicting the authenticity of a review and investigating metadata-based features that integrate SBERT-based similarity measures for predicting review helpfulness. Drawing on the theories of informative richness, linguistic analysis, and consumer behavior, the following hypotheses are proposed:

H1: Linguistic and stylistic-based features, such as authenticity score, can significantly predict a review's credibility.
H2: The Sentiment polarity and score of a review are relevant to its credibility.
H3: Reviewer credibility attributes are significantly associated with review helpfulness.
H4: The informativeness of review content, measured by SBERT-based similarity to product metadata, is positively associated with review helpfulness.

To synthesize these hypotheses, this research employed various NLP and ML-based computational methods:

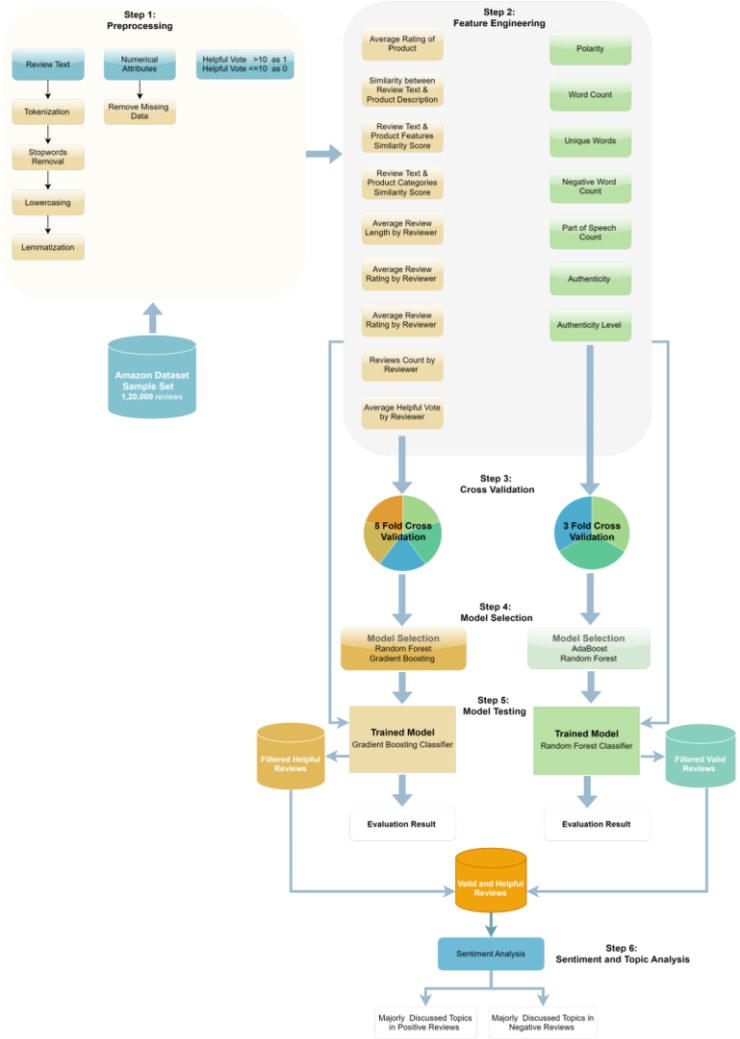
(1) Linguistic Analysis: Features derived from the linguistic aspects of the review, such as the genuineness score, are rooted in linguistic theory, which posits that specific patterns inconsistent with human manner of expression can be utilized to distinguish fake and genuine reviews.
(2) SBERT and cosine similarity: SBERT can be utilized to generate meaningful vectors for review text and product metadata (Kühl et al., 2020; Sayeed et al., 2023). The informativeness of a review regarding a product can be computed using the cosine similarity between the review and product metadata vectors. Cosine similarity in vector space modeling is a foundational method to measure the angle between two encodings, ranging between 1 and -1, with 0 being no semantic overlap (Jyoti & Singh, 2015; Park et al., 2020).
(3) Machine Learning Algorithms: Tree-based algorithms, such as Random Forest, Gradient Boosting, and AdaBoost, are employed to be trained on the training set and then predict outcomes. The models

were chosen due to their interpretability and proven effectiveness in classification tasks (Abd & Hussein, 2024; Bilal & Almazroi, 2023; Salminen et al., 2022).

4. Methodology

Primarily, the methodology encompasses three primary components: helpfulness classification, fake reviews classification, and sentiment-driven topic extraction with subsequent analysis. As illustrated in Figure 1, the process begins with comprehensive data pre-processing with a focus on retaining only semantically relevant terms. This is preceded by classifier development, encompassing the training for both classification models for helpfulness detection and fake reviews identification individually. This is followed by a rigorous evaluation of the model's performance through standard performance metrics. The relation between the performance of the models and the features is also analyzed to assess their contribution to predictive efficacy.

Figure 1. Methodology



Note: The methodology of the proposed two-step framework includes preprocessing, feature engineering, cross-validation, classifier training, and their evaluation. Finally, the classifiers are used for filtering and retaining only helpful and valid reviews, and sentiment-based topic analysis is performed.

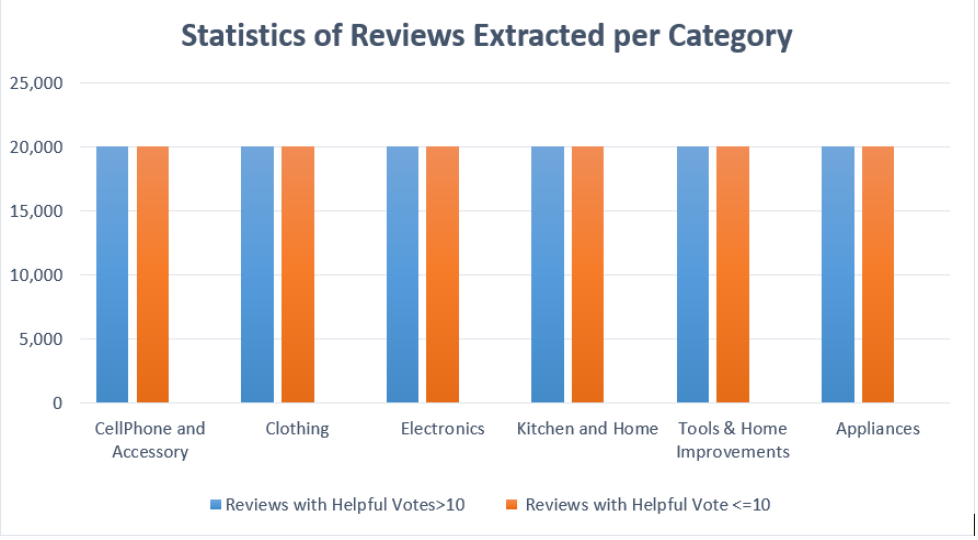
The trained classifiers generate probabilistic estimates for review helpfulness and its authenticity, allowing the systematic removal of unhelpful and fraudulent reviews from the dataset. This filtering ensures the usage of only helpful and real reviews for subsequent analysis. Finally, insights regarding product feature perception are obtained through conducting sentiment-based topic analysis on the reviews regarding a product. These insights facilitate the understanding of positively evaluated merchandise features, thereby allowing evidence-based decision-making grounded in helpful and authentic feedback.

The two-stage integrated methodology ensures data quality through multi-stage validation while retaining the semantic richness required for insightful topic extraction and sentiment analysis.

4.1 Data Collection

Some attributes from the publicly available Amazon Dataset (Hou et al., 2024) have been selected for the helpful reviews classification model. The extracted attributes for each product include 'Product ID', 'Product Title', 'Product Link', 'Product Rating', 'Product Price', 'Product Description', and 'Reviews'. Review attributes include 'Reviewer Name', 'Review Rating ', 'Review Body', and 'Review Helpfulness Score'.

Figure 2. Statistics of Reviews in the Dataset



Note: The number of samples is given along the y-axis, and the name of the product category is given along the x-axis. In total, 120,000 reviews were collected, with half of them having a helpful vote of> 10 and half of them with a helpful vote of <= 10.

The statistics of extracted reviews for the six general categories are shown in Figure 2. A review is considered accommodating if the number of accommodating votes exceeds 10; otherwise, it is labeled as unhelpful. The label for helpful reviews is set to 1, and the label for unhelpful reviews is set to 0. We extracted 40,000 reviews from e-commerce platforms for each category and balanced them between helpful and unhelpful votes.

For the determination of false reviews, another dataset (Hussain et al., 2020) was pre-processed from the Amazon dataset. The dataset comprises a large-scale collection of Amazon reviews, with data from a

single category, "Toys and Games," used to identify fake or genuine reviews. The labeling procedure involved manual scrutiny, where commentators or annotators evaluated the reviews based on explicit features that indicated whether a review was spam or genuine. The data is stored in a CSV file.

4.2 Preprocessing

These preprocessing techniques enhance the quality of text representations, minimize text noise, and transform the script into a meaningful design (Alasadi & Bhaya, 2017). For the review text content, we follow a few key steps: tokenization (Duong & Nguyen-Thi, 2021), lowercasing, removing stop words, and lemmatizing (Işik & Dağ, 2020) the words to ensure uniformity and reduce noise in the text. Missing values in the dataset need to be handled appropriately before feature engineering. In this context, the descriptions of most products were missing in the raw dataset. Hence, they were replaced by empty strings instead to ensure structural consistency. After the above preprocessing steps, we make the tokens a single string for each review, providing a semantically cleaned textual representation. After that, the resulting string shows the preprocessing text used for further analysis.

We process the "helpful vote count" feature to create a new label. Reviews with as many as 10 helpful votes are labeled 1, meaning they are considered beneficial. Reviews with 10 or fewer helpful votes are labeled as 0, indicating that they are not considered helpful. This new labeling helps the model distinguish between helpful and less helpful reviews.

4.3 Feature Extraction

The features extracted and engineered for a helpful review recognition are primarily concerned with the ability to be informative and the reviewer's trustworthiness. In contrast, characteristics extracted for false review recognition focus on the stylistic content of the review. The details regarding each classifier's feature engineering process have been provided under their respective heading below:

4.3.1 Features For Identifying Helpfulness of Reviews

Aspects of review data generally include ratings and verified purchase status, which provide a primary understanding of the product's performance and customer feedback. Still, they do not entirely gather the customer's liking. The resulting features can provide more meaningful information when we augment these basic features with the similarity between review text and numerous product details to measure how closely the review matches the product information.

The metadata-based review features include: rating of review, whether the purchase was verified against which the review was made, average rating of a product across all users, the total number of reviews for a product, and the length of the review. These features capture the product-level signals through average_rating, rating number, and review-level signals through rating, verified_purchase, and review_length, together providing a comprehensive view of product quality and review reliability.

To quantify the informativeness of a review, we measure its semantic alignment with product-level metadata. As product-level metadata, we include product description, product features, and categories. The review content-based features, similarity_text_categories, similarity_text_features, and similarity_text_description, are generated by encoding the review sentence by sentence and then finding similarities with the respective product information. SBERT examines each word from the entire sentence by processing the text bi-directionally, capturing both syntactic and semantic relations among words (Kühl et al., 2020; Sayeed et al., 2023). Doing this enables a comprehensive understanding of the text, which allows us to convert the sentences into embeddings that accurately represent the complete meaning of the text. This improves our analysis with semantics, which is important for accurate similarity calculations.

In our research framework, we utilize cosine similarity to evaluate the similarity between two sentences, treating each sentence as a vector due to its efficiency in handling sparse data and scale invariance. Cosine similarity usage has been supported by research on semantic matching tasks. Leveraging BERT's pre-trained language models, we incorporate detailed contextual information relevant to the input text, thereby enriching our vectors with a deep semantic understanding (Singh et al., 2022). To compute the resemblance, we use the following formula that provides the cosine of the angle θ between two vectors A and B, which is known as cosine similarity:

$$Cosine\ Similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

where n is the total number of categories, the similarity is the resemblance score between the input script A of category i details and the input script B of category i. The range of cosine similarity values ranges between -1 and 1, where represents that the vectors are in the same direction. Still, they are perfectly similar; -1 represents vectors in opposite directions, and 0 means no similarity. For a review, its similarity_text_features attribute has a larger value if the review text contains product feature details.

The review's usefulness is also dependent on the reviewer's behavior patterns. We also incorporate the reviewer's average review length into our analysis, as a reviewer with a greater average review length is more likely to provide a helpful review, as shown in Equation 2.

$$Average\ Review\ Length = \frac{\sum_{i=1}^{N_r} L_i}{N_r}, \quad (2)$$

where L_i denotes the length of i^{th} review written by the reviewer, and N_r is the total number of reviews by the reviewer.

The review writer's average review score and review count indicate the reviewer's rating behavior pattern. A spam reviewer may provide an average review rating of mostly 0. The average rating is the mean value of all reviewer ratings as defined in Equation 3.

$$Average\ Review\ Rating = \frac{\sum_{i=1}^{N_r} R_i}{N_r}, \quad (3)$$

where R_i denotes the rating assigned by the reviewer in the i-th review, and N_r is the total number of reviews by the reviewer. While review ratings on a scale of 1-5 are inherently ordinal, they are treated as approximately continuous in this study to calculate the average review rating. This is supported by literature that demonstrates that parametric statistics remain robust when applied to 5-point scales (Norman, 2010). Averaged ratings have been previously adopted in review analysis and recommendation system literature to model rating behavior (L. Chen et al., 2015).

Another feature we have utilized is the mean number of helpful votes a reviewer's reviews receive, which indicates the perceived usefulness of a reviewer's contributions to the community. A consistently low average might indicate less impactful or potentially spammy reviews. It is given by:

$$Average\ Helpfulness\ Votes = \frac{\sum_{i=1}^{N_r} H_i}{N_r}, \quad (4)$$

where H_i denotes the helpfulness votes received by the i-th review written by the reviewer, and N_r is the total number of reviews by the reviewer.

4.3.2 Features For Identifying Fake Reviews

The features that should be utilized to identify fake reviews are more geared towards the writing style and nature of words in the text. Hence, new features are generated from the review content to encode the stylistic features of an analysis. The details regarding newly generated features are given below:

Three aspects of features have been extracted from the review text: sentiment-based, word-based, and parts-of-speech-based. Sentiment analysis-based features helped identify reviews and assess people's thoughts and feedback extracted. We have selected the TextBlob model for sentiment analysis due to its lower computational complexity while maintaining quality. Word-based features include the total number of lexical items in feedback, the sum of different lexical units in each feedback, and the sum of words with negative sentiment. Parts of speech features comprised the sum of nouns, pronouns, verbs, auxiliary verbs, articles, conjugations, propositions, adjectives, and adverbs in the feedback.

Genuineness and a Threshold T are two features, defined by Equations 5 and 6, respectively, that are then derived using the above features. Individual reviews indicate a higher genuine score containing more pronouns and spontaneous write-ups with a less defensive tone. Threshold T represents the use of more articles, and the preposition count suggests a formal and informational tone in the message. A higher ratio of pronouns, auxiliary verbs, conjunctions, negations, and adverbs, as well as a higher negative word count, compared to articles and prepositions, represents a more informal tone.

$$Genuineness = \frac{pronoun\ count + unique\ word\ count - negative\ word\ count}{total\ word\ count}, \quad (5)$$

where $pronoun\ count$ denotes number of pronouns, $unique\ word\ count$ denotes number of unique word in the review, $negative\ word\ count$ denotes the number of negative words in review, and $total\ word\ count$ denotes the total number of words in the review.

$$T = \frac{art\ count + prep\ count - pro\ count - aux\ count - conj\ count - adv\ count - neg\ word\ count}{word\ count}, (6)$$

where $art\ count$ denotes number of articles, $prep\ count$ denotes number of prepositions, $pro\ count$ denotes number of pronouns, $aux\ count$ denotes number of auxiliary verbs, $conj\ count$ represents number of conjugations, $adv\ count$ denotes number of adverbs, and $neg\ word\ count$ denotes number of negative word in a review.

4.4 Experimental Setup

Specifications regarding the hardware and software used are listed in Table 2.

Table 2. Experimental Setup

Hardware and Software	System Specifications
Processor	Intel Xeon CPU with 2 vCPUs(Virtual CPUs)
GPU	NVIDIA K80 12 GB
RAM	13 GB
Environment	Python: 3.8
Storage	100 GB SSD cloud storage

Note: The experiments were conducted in a Google Colab Free-tier Standard Environment.

We conducted Augmented Dickey-Fuller (ADF) (Dickey & Fuller, 1979) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) (Kwiatkowski et al., 1992) tests as a precautionary check to assess the presence of unit-root-like behavior in the observation sequence. We applied the KPSS test, which has a null hypothesis of stationarity, and the ADF test, which has a null hypothesis of non-stationarity. The unit root testing was conducted using Python's statsmodels library.

In time series contexts, research has demonstrated that spurious relationships may emerge with almost non-stationary processes (Cheng et al., 2021, 2022), with variables of auto regressive dynamics (Wong & Pham, 2022a, 2022b, 2023a, 2023b), and even among stationary series under certain conditions(Wong et al., 2024; Wong & Pham, 2025; Wong & Yue, 2024a, 2024b). These studies illustrate the inferential limitations of standard parametric t-tests and F-tests when observations are not independent in time series contexts. To address these concerns, we applied the following criteria: For the KPSS test, we defined p>=0.05 as stationarity, and a series is flagged as nearly non-stationary if either the p-value fell between 0.05 and 0.10, or if the test statistic is very close to the 5% critical value. For the ADF test, we classified stationarity using the standard hypothesis testing criterion of $p < 0.05$, and a series was labeled nearly non-

stationary if its p-value fell between 0.05 and 0.10, or if its test statistic was within 1.0 of the 5% critical value.

Table 3. Unit Root Test Results (ADF Test) for the dataset used for helpful reviews classification

Variable	% Stationary (ADF)	% Nearly Non-Stationary (ADF)
Rating	86.4	18.2
helpful_vote	85.4	16.4
average_rating	85.4	17.0
rating_number	83.8	13.9
similarity_text_description	85.4	21.9
similarity_text_features	88.8	15.1
similarity_text_categories	87.1	20.7
similarity_text_details	88.3	18.9
sentiment_scores	86.6	15.3
review_length	79.3	28.4

Note: The Table indicates the ADF test result with percentages indicating the entity-level series classification as stationary ($p < 0.05$) or nearly non-stationary. Nearly non-stationary series are those with p-values between 0.05 and 0.10 or test statistics close to critical values, reflecting the borderline unit-root behavior.

Table 4. Unit Root Test Results (KPSS Test) for the dataset used for helpful reviews classification

Variable	% Stationary (KPSS)	% Nearly Non-Stationary (KPSS)
Rating	87.1	27.9
helpful_vote	86.6	26.7
average_rating	76.4	30.4
rating_number	87.2	30.6
similarity_text_description	72.8	33.3
similarity_text_features	81.0	35.0
similarity_text_categories	80.4	33.8
similarity_text_details	81.0	36.4
sentiment_scores	87.9	29.9
review_length	71.9	35.9

Note: The Table indicates the KPSS test result with Percentages indicating the entity-level series classification as stationary($p>=0.05$) or nearly non-stationary. Nearly non-stationary series are those with test statistics $p \in [0.05, 0.10]$ or test statistics close to the critical value or p-values close to thresholds, reflecting borderline stationarity.

The results in Table 3 and Table 4 show that, using a majority-based approach for entity-level series classification, across all variables, more than 60% of the series are classified as stationary under both ADF and KPSS tests, respectively. The proportion of nearly non-stationary cases remains limited and non-dominant. We also performed Data Generating Process (DGP) persistence analysis and found that the entity-level AR(1) coefficients (ρ) for all variables were less than 0.8 for a substantial fraction of entity processes (Cheng et al., 2021, 2022). Results show that rating_number is the only variable with high persistence, while all other variables are low-risk for regression-based analysis. Random Forest and Gradient Boosting are tree-based models that construct predictions through recursive partitioning based on feature-outcome relation observed in the data, without parametric assumptions about trends, temporal dependence, or independent observations. They are designed to capture complex, nonlinear interactions

without the need to be transformed to a stationary state. Our research objective is cross-sectional classification of whether an individual review is helpful based on its characteristics rather than forecasting how helpfulness evolves over time. The relation between review features describing informational content, such as review length, rating of a review, similarity scores, and its helpfulness, reflects stable informational value rather than a temporal pattern. Furthermore, to ensure that feature importance is not biased by trending data or spurious correlations, we evaluated feature permutation importance with statistical validation. Each feature was randomly permuted over multiple iterations, and the resulting decrease in model performance was tested against the null hypothesis that no predictive signal (H_0: $\mu = 0$) using a one-sample t-test with Bonferroni correction for multiple comparisons. When a feature is randomly shuffled, any spurious temporal patterns or coincidental trends are broken, while genuine feature-outcome relationships are preserved. If a feature's importance is derived from spurious correlation, permutation would not significantly degrade model performance. Conversely, features showing statistically significant performance degradation demonstrate genuine predictive relationships robust to the limited non-stationarity. All features showed statistically significant performance degradation upon permutation ($p < 0.001$), confirming genuine predictive value. The results are presented in detail in the results and discussion section. While rating number displays high persistence, which could potentially lead to spurious correlations in classical regression frameworks, our approach differs by using a supervised machine learning approach, which is supported by prior research on review helpfulness prediction commonly incorporating rating-related metadata as input features. Including product rating features has shown valid predictive performance in Random Forest baselines and other classifiers (Bilal et al., 2019; Zhou & Yang, 2019). Aggregated features, such as reviewer-level features in the helpful review dataset and review-text-derived features in the fake review dataset, are computed after data collection as summary statistics and do not represent time-based observations. These aggregated features do not exhibit temporal dependencies and therefore do not require stationarity assessment.

4.5.Model Selection and Evaluation Parameters

We evaluated various classifiers after training to identify a suitable classifier for the complex tasks of classifying helpful reviews and fake reviews. The details of these algorithms are given below:

For the helpful reviews classifier training, the dataset has been divided into a testing set and a training set with an 80:20 ratio to compute the model's performance. Regarding the details of the data preparation strategy for identifying patterns in fake reviews, we have used a 60-40 split.

For classifying product reviews into two categories, "helpful" and "unhelpful", two different models, Random Forest and Gradient Boosting, have been trained with cross-validation over a 5-fold approach. The Random-Forest classifier can effectively control extreme data dimensionality as well as multilinearity, being quick and insensitive to overfitting (Bilal & Almazroi, 2023; Kühl et al., 2020). Random Forest is a technique that combines numerous decision trees to classify large datasets accurately.

Let $\hat{C}_b(x)$ be the class prediction of the b_{th} decision tree in the Random Forest (Breiman, 2001). The overall Random Forest classifier prediction, $\hat{C}_{rf}^{B}(x)$, is defines as follows:

$$\hat{C}_{rf}^{B}(x) = majority\ vote\ \{\hat{C}_b(x)\}_{\mathrm{b}=1}^{B}, \tag{7}$$

where $\hat{C}_b^{\square}(x)$ is the predicted class label from the b_{th} tree, B is the total number of trees in the forest, and x represents the input feature vector. The final classifier prediction is determined by the majority vote among all tree predictions.

On the other hand, Gradient Boosting focuses more on minimizing error than Random Forest's averaging approach. A minor learning degree is applied to scale the contribution of each tree, allowing for fine-tuning the corrections (Friedman, 2001). The Gradient Boosting prediction function, $\hat{C}_{GB}^{B}(x)$, is given below:

$$\hat{C}_{GB}^{B}(x) = sign\left(\sum_{b=1}^{B} \eta\ .h_b(x)\right), \tag{8}$$

where $h_b(x)$ is the prediction from the b-th tree, η is the learning rate (a small scalar that controls how much each tree has contributed to the concluding expectation), B is the total number of boosting iterations, x is the input vector, and a $sign$ function maps the prediction to the final class label in binary classification.

The classification of fake reviews involves detecting reviews that are misleading or falsified. In classification, machine learning models like AdaBoost (Adaptive Boosting) and Random Forest have been trained and cross-validated using a 3-fold approach. AdaBoost, by combining numerous weak learners, typically decision trees, creates a robust learner where each succeeding model emphasizes the errors made by the prior ones (Freund & Schapire, 1997). The concluding sturdy classifier $H(x)$, is a biased combination of all other feeble classifiers:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t\ .h_t(x)\right), \tag{9}$$

where $H(x)$ is the concluding classifier, $h_t(x)$ is the prediction output of the t^{th} weak classifier, α_t is the weight of the t^{th} feeble classifier, T is the overall sum of iterations, and x is the input feature vector, and a $sign$ determines the final output class label in binary classification

All classifiers in this work have their hyperparameters determined through the hyperparameter tuning technique of grid search (Xu et al., 2024). Grid search is a common technique for hyperparameter tuning that tests all possible combinations of detailed hyperparameters and selects the best-performing set, although it can be computationally expensive (Campos et al., 2025). The method ensures an additional accurate approximation of the models' performance by applying a 5-fold cross-validation for helpfulness classification and a 3-fold cross-validation for fake review classification.

For supervised learning, a structured manner of analyzing a model's performance is through its confusion matrix, which allows comparison between predicted and actual outcomes. A confusion matrix is based on four measures: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). True Positives represent instances where the model correctly predicts a positive class label, while True Negatives represent instances where the model correctly predicts the negative class label. A False Positive is the case when the model incorrectly predicts a sample to be from the positive class, while a False Negative is when the model incorrectly labels a sample to be from the negative class (Tharwat, 2021).

After the training of the model, we analyzed its performance using test data and obtained the following evaluation scores:

(1) Accuracy represents the ratio of correctly classified occurrences to the overall occurrences, and is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

(2) The Precision represents the percentage of true positive instances among the overall instances, and is given by:

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

where TP is the number of true positives, and FP is the number of false positives.

(3) The Recall symbolizes the percentage of true positive instances out of genuine positive ones, and is defined as:

$$Recall = \frac{TP}{TP + FN}, \tag{12}$$

where TP is the number of true positives, and FN is the number of false negatives.

(4) The F1-Score, denoted as $F1$, offers a stable measure of a model's performance by harmonic means of precision and recall, as follows:

$$F1 = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}, \tag{13}$$

where Precision and Recall are defined in Equations 11 and 12, respectively.

(5) The Area Under the Curve, AUC, measures the classifier's ability to distinguish between the two classes. It is the integral of the True Positive Rate (TPR) as a function of False Positive Rate (FPR) across all possible thresholds. The TPR is the same as recall, and the FPR is given by the following formula:

$$FPR = \frac{FP}{FP + FN}, \tag{14}$$

where FP is the number of false positives, and FN is the number of false negatives.

Finally, the formula for Area Under the curve is given by:

$$AUC = \int_0^1 \mathrm{TPR}\left(\mathrm{FPR}^{-1}(\mathrm{x})\right) \mathrm{dx}, \tag{15}$$

where x∈[0,1] is a continuous variable that denotes a false positive rate value, $\mathrm{FPR}^{-1}(\mathrm{x})$ is the threshold that produces a given FPR of x, $\mathrm{TPR}\left(\mathrm{FPR}^{-1}(\mathrm{x})\right)$, denotes the True Positive Rate obtained at threshold that yields a False Positive Rate of x.

(6) Feature importances are computed by the average reduction in impurity for the tree-based algorithms and by the weighted contribution of weak learners for the boosting algorithms. For model interpretation, it is crucial to identify which variables contribute to the predictive capability of the model. For Random Forest, the feature importance FI_{rf}^{j}, of each feature j is computed by (Pedregosa et al., 2011):

$$FI_{rf}^{j} = \frac{1}{T}\sum_{t=1}^{T}\sum_{n \in N_{j,t}} \frac{w_n}{w_{total}} \Delta i(n), \tag{16}$$

where FI_{rf}^{j} denotes the importance of feature j, T is the total number of trees, $N_{j,t}$ are the nodes where feature j was used in tree t, w_n is the weighted number of samples at node n, w_{total} is the total number of weighted samples in the entire training dataset and $\Delta i(n)$ represents the impurity decrease caused by the split.

Gradient Boosting uses the total gain from the splits across all trees to compute Feature Importances. The gain measures improvement in the loss function due to a split. The formula for calculating feature importance for Gradient Boosting for feature j, FI_{GB}^{j} , is given in Equation 17 (Chen & Guestrin, 2016).

$$FI_{GB}^{j} = \frac{\sum_{t=1}^{T}\sum_{n \in N_{j,t}} Gain(n)}{\sum_{t=1}^{T}\sum_{n \in N_t} Gain(n)}, \tag{17}$$

where $Gain(n)$ represents improvement in the loss function upon the split based on feature j at node n, $N_{j,t}$ represents nodes where feature j was used in tree t , T is the total number of trees, and N_t represents all nodes in tree t.

Feature importance in the AdaBoost model using the Scikit-Learn library in Python can be intuitively written as given in Equation 18. For each weak learner t, we take its weight α_t and multiply it by the impurity decrease measure contributed by that feature in that tree (Friedman, 2001). Feature importances for the Adaboost model for feature j, FI_{AB}^{j}, are computed by the following formula:

$$FI_{AB}^{j} = \sum_{t=1}^{\top} \alpha_t \, Usage_{j,t}, \tag{18}$$

where α_t represents the contribution weight of the weak learner t, $Usage_{j,t}$ represents whether feature j was used in the weak learner t, and $\top$ is the total number of trees.

(7) Permutation Importances have been computed using a model-agnostic method to quantify the extent of a model's dependence on predictive power on each feature individually. It is different than feature importance scores because it measures the change in predictive performance of the model upon random shuffling of the features. Upon random shuffling, the relationship between the feature and the target feature is disrupted, which enables the quantification of the dependence of a model on that feature (Breiman, 2001; Fisher et al., 2019). The permutation importance PI_j of a feature j is computed as:

$$PI_j = \frac{1}{R}\sum_{r=1}^{R} \left[M(D) - M\left(D^{\pi(j,r)}\right)\right], \tag{19}$$

where, $M(D)$ is the models performance score on Dataset D, π denotes a permutation operator, $\pi(j,r)$ represents random permutation applied to feature j in the r^{th} reptition, $M\left(D^{\pi(j,r)}\right)$ is the model's performance on dataset with randomly permuting feature j in the r^{th} repetition while all other features remain unchanged, and R is the total number of repititions. Higher scores of PI_j implies greater dependence of the model on feature j.

5. Results and Discussion

This study addresses the current research on evaluating user-generated content quality, product/service enhancement, and marketing strategies using text mining tools in e-commerce practices, which have replaced consumer surveys (Li et al., 2022). Online reviews can significantly impact a consumer's purchase decision. Thus, identifying the reviews that receive the most attention will benefit retailers in improving product sales by enhancing aspects of the product discussed (Li et al., 2024). In this study, we attempt to analyze the role of review content, its author, and product characteristics in implementing an automatic labeling of review helpfulness. This study focuses on the classification of review helpfulness for all customers by suggesting new features that have received little attention in the past. The results of our study demonstrate that review content alone is insufficient for classifying a review as helpful. Still, it is necessary to understand the informativeness of the review content. An informative review is more likely to be rated as helpful by customers.

A classification report, provided in Tables 6 and 8, summarizes the evaluation metrics, offering a thorough understanding of the model's workings for helpful review classification. The Random Forest achieved 94% accuracy with a 98% Area Under the Curve in predicting the helpfulness of a review. The best

parameters obtained for Random Forest are 'max_depth': None, 'n_estimators': 150, and 'minimum_samples_split':2 using grid search with n_estimators: [50,100,150], max_depth : [None, 3, 5, 7], and min_samples_split : [2, 5, 10]. The confusion matrix is also shown for Random Forest in Table 5.

Table 5. Confusion Matrix for Random Forest Classifier for helpful reviews classification

	Predicted 0	**Predicted 1**
Actual 0	18,193	1,701
Actual 1	1,082	27,024

Note: The above confusion Matrix indicates strong model performance with significantly high true positives and true negatives with minimal incorrect predictions. Class 0 stands for not helpful, and Class 1 represents helpful.

The confusion matrix shown in Table 5 demonstrates that the Random Forest classifier performs strongly in distinguishing between helpful and unhelpful reviews. A high number of correct predictions for Class 1(helpful) of 27,024 indicates the model correctly identifies most of the helpful reviews. Similarly, large correct predictions for the Class 0 (unhelpful class) show effective filtering of the unhelpful reviews. False positives indicate occasional overestimation of helpfulness, but the impact is minor in comparison to the overall dataset.

Table 6. Classification Report for Random Forest Classifier for helpful

	Precision	**Recall**	**F1-Score**	
0	0.94	0.91	0.93	
1	0.94	0.96	0.95	
			Accuracy:	0.945

Note: The table presents the evaluation metrics of the Random Forest Classifier. The recall for class 1 is better than class 0, indicating that the model is better at detecting helpful reviews than unhelpful ones. Precision is the same for both classes. Finally, the F1-score for both classes is above 90%, demonstrating that the model is well-balanced and effective. Evaluation metrics were calculated by applying the methodology defined in Equations 10 - 13.

The Random Forest Classifier shown in Table 6 demonstrates high reliability with a 94.5% accuracy score. High recall for helpful reviews means that the majority of the helpful reviews are retained by the classifier. The ability to correctly filter out non-helpful reviews is 91.4%, demonstrating its effectiveness to reduce noise and improve information quality on e-commerce platforms. It provides balanced precision and recall for the helpful class, reflecting that the model avoids bias either toward predicting helpfulness or being overly negative.

The gradient boosting algorithm also achieved 94% accuracy and 98% AUC score in predicting the helpfulness of a review. The best parameters found using hyperparameter tuning are learning rate: 0.1, maximum_depth: 7, and 'n_estimators': 150 using a parameter grid of 'n_estimators': [50,100,150] and 'max_depth': [3,5,7]. The confusion matrix is also shown for the Gradient Boosting Algorithm in Table 7.

Table 7. Confusion Matrix of Gradient Boosting Classifier for helpful reviews classification

	Predicted 0	**Predicted 1**
Actual 0	18,478	1,644
Actual 1	1,045	26,833

Note: The above confusion Matrix highlights its strong ability to predict helpful reviews (Class 1) with high accuracy and minimal misclassifications. Gradient Boosting demonstrates slightly better performance for helpful reviews than Random Forest.

The confusion matrix of the Gradient Boosting classifier, shown in Table 7, shows a high number of correct classifications for both helpful and unhelpful classes, as demonstrated by the high true positives of 26,833 and high true negatives of 18,478. Lower false negatives and false positives highlight the minimal underestimation and overestimation of helpfulness.

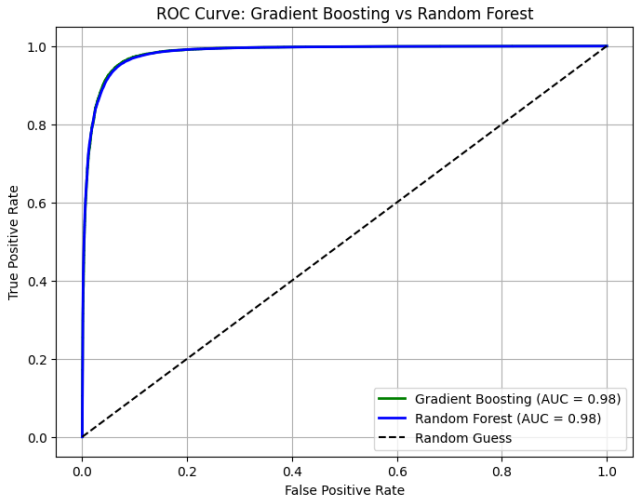
Table 8. Classification Report for Gradient Boosting Classifier for helpful reviews classification

	Precision	**Recall**	**F1-Score**	
0	0.94	0.92	0.93	
1	0.94	0.96	0.95	
			Accuracy:	0.94

Note: The above classification report shows consistent balanced performance for both classes, but better recall and F1-score for class 1 (helpful reviews). These metrics indicate the model's effectiveness in review quality filtering in the e-commerce domain. Evaluation metrics were calculated by applying the methodology defined in Equations 10 - 13.

As the results shown in Table 8, the overall accuracy of 94% confirms the reliability of the model. Gradient Boosting provides balanced precision and recall classification with fewer false helpful reviews shown and fewer helpful reviews missed. The model offers dependable filtering, which can reduce noise and improve information quality.

Figure 3. ROC curve comparison between Random Forest and Gradient Boosting classifiers for Helpful reviews prediction



Note: The number of false positives is given along the x-axis, while the true positive rate is given along the y-axis. The diagonal dashed line denotes the random guess baseline. Both models provide the same AUC. AUC scores were calculated by applying the methodology defined in Equations 14 - 15.

The Receiver Operating Characteristic curve is shown in Figure 3 to compare the performance of Gradient Boosting and Random Forest for predicting helpful reviews. Both models achieved an identical high AUC

of 0.98, indicating high discriminative power. Random offers a practical advantage by achieving a slightly higher number of correctly classified helpful reviews, meaning it retains more genuinely useful content. This is crucial in consumer-centric applications where retaining informative reviews is more valuable than identifying unhelpful ones. Our results demonstrate how a Random Forest classifier can be trained to achieve excellent performance in predicting whether a review is helpful.

To assess the statistical significance of the classifiers, permutation importance μ is calculated for over 30 iterations. Statistical validation was conducted through a one-sample t-test against the null hypothesis H_0: $\mu = 0$, which denotes that a feature has no importance. The p-values were adjusted using Bonferroni correction, and features were only considered significant at a threshold of the significance level $\alpha = 0.05$. Confidence intervals of 95% were computed as $\mu \pm 1.96\sigma$, where σ is the standard deviation. The permutation importance scores for the Random Forest classifier are shown in Table 9.

Table 9. Permutation Importance with one-sample t-test scores for Random Forest Classifier for helpful reviews classification

Feature	Mean Importance	95% CI (Lower–Upper)	Adjusted p-value
reviews_count	0.241***	0.239 – 0.242	p < 0.001
avg_review_length	0.198***	0.197 – 0.199	p < 0.001
rating_number	0.075***	0.074 – 0.076	p < 0.001
similarity_text_categories	0.064***	0.064 – 0.065	p < 0.001
review_length	0.043***	0.043 – 0.044	p < 0.001
avg_rating	0.042***	0.042 – 0.043	p < 0.001
similarity_text_description	0.042***	0.042 – 0.043	p < 0.001
average_rating	0.040***	0.039 – 0.040	p < 0.001
similarity_text_features	0.020***	0.019 – 0.020	p < 0.001
verified_purchase	0.009***	0.009 – 0.009	p < 0.001
rating	0.004***	0.003 – 0.004	p < 0.001

Note: *** indicates statistical significance at $p < 0.001$ after Bonferroni correction. The table reports the mean decrease in feature importances when the feature is permuted with 95 % confidence intervals (CI). The name of the feature is provided along with its mean permutation importances, its upper and lower bounds for the confidence interval calculated as $\mu \pm 1.96\sigma$. The p-value, i.e., the probability of observing the obtained importance value under the null hypothesis, adjusted using the Bonferroni method, is provided in the table. Statistical significance is true when the adjusted p-value is less than 0.05. Permutation Importance scores were calculated by applying the methodology defined in Equation 19.

It was observed in Table 9 that all features are statistically significant, meaning that each feature contributes non-random information to the Random Forest model. The reviews count, and average review length are the strongest predictors with high importance. The similarity score between review text and categories, review length, average rating by reviewer, similarity between review text and product description, and average rating by reviewer are medium predictors. Hence, our study demonstrates how review text can be transformed into more significant attributes using the deep learning-based encoding technique of SBERT and simple techniques such as cosine similarity. Also, review metadata, such as

rating and verified purchase, contributed marginally to differentiate helpful reviews from unhelpful ones. The permutation importance scores for the Random Forest classifier are shown in Table 10.

Table 10. Permutation Importance with one-sample t-test scores for Gradient Boosting Classifier for helpful reviews classification

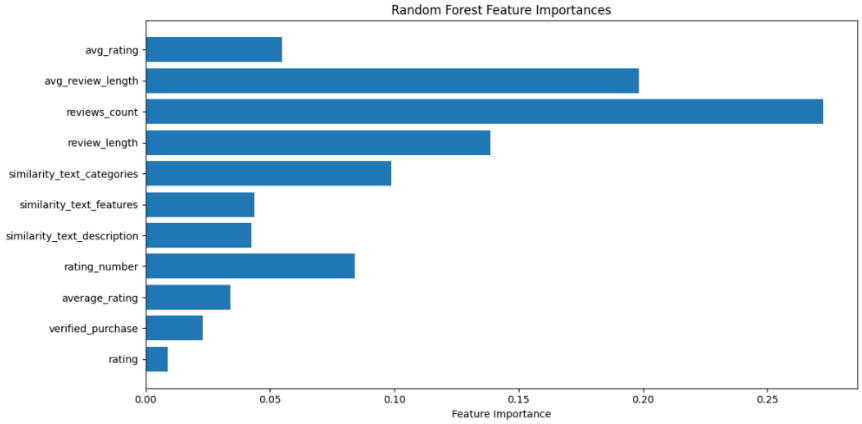
Feature	Importance Mean	95% CI (Lower–Upper)	Adjusted p-value
reviews_count	0.2406***	0.2390 – 0.2421	$p < 0.001$
avg_review_length	0.1980***	0.1968 – 0.1992	$p < 0.001$
rating_number	0.0751***	0.0744 – 0.0758	$p < 0.001$
similarity_text_categories	0.0644***	0.0637 – 0.0651	$p < 0.001$
review_length	0.0433***	0.0427 – 0.0439	$p < 0.001$
avg_rating	0.0425***	0.0419 – 0.0430	$p < 0.001$
similarity_text_description	0.0422***	0.0416 – 0.0429	$p < 0.001$
average_rating	0.0395***	0.0390 – 0.0401	$p < 0.001$
similarity_text_features	0.0197***	0.0193 – 0.0200	$p < 0.001$
verified_purchase	0.0091***	0.0089 – 0.0093	$p < 0.001$
rating	0.0036***	0.0034 – 0.0037	$p < 0.001$

Note: *** indicates statistical significance at $p < 0.001$ after Bonferroni correction. The table reports mean permutation importances of the features in descending order for the gradient boosting classifier. Their upper and lower bound is given for a confidence interval calculated as $\mu \pm 1.96\sigma$. The p-value, i.e., the probability of observing the obtained importance value under the null hypothesis adjusted using the Bonferroni method, is provided in the table. Statistical significance is true when the adjusted p-value is less than 0.05. Permutation Importance scores were calculated by applying the methodology defined in Equation 19.

As the results shown in Table 10, the mean permutation importances analysis reveals that all the features are statistically significant for the gradient boosting classifier. Reviewer history dominates the prediction, with average reviews count and average review length contributing strongly. The similarity between review text and categories also contributes significantly, implying that product metadata is a significant cue for predicting a review as helpful. Review metadata, such as verified purchase and rating, has a modest impact.

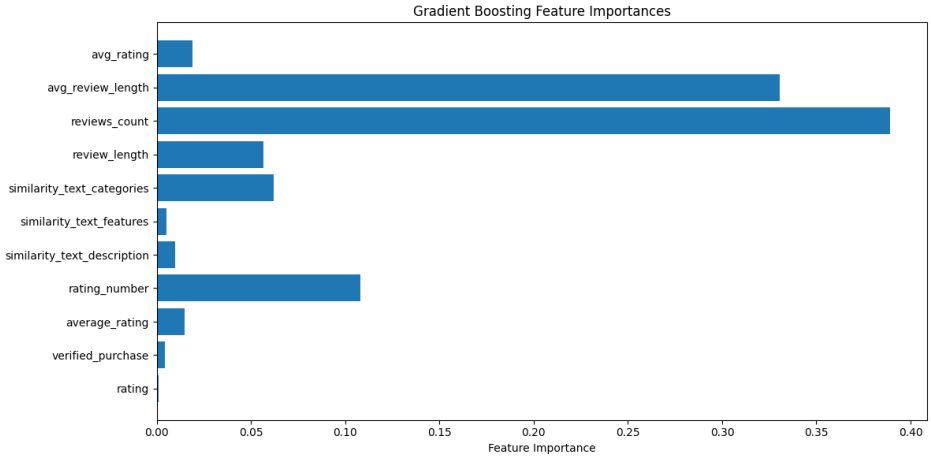
To gain further insight into the performance of the models, their feature importances are displayed in Figures 4 and 5. The Figure shows feature importance on the x-axis and feature name on the y-axis. Both figures show that reviews_count and average_review_length are the most dominant contributors to the model. This suggests that the reviewer's behavioral history is related to information richness and serves as a reliable indicator of helpfulness. Random Forest distributes its importance among diverse features while Gradient Boosting dominantly relies on the identification of strong features alone.

Figure 4. Feature Rankings for Random Forest Classifier for helpful reviews classification



Note: The Figure shows feature importance for the Random Forest model on the x-axis and feature name on the y-axis. Feature importance scores were calculated by applying the methodology defined in Equation 16.

Figure 5. Feature Rankings for Gradient Boosting Classifier for helpful reviews classification



Note: The Figure shows feature importance for the Gradient Boosting model on the x-axis and feature name on the y-axis. Feature importance scores were calculated by applying the methodology defined in Equation 17.

In conclusion, both Figure 4 and Figure 5 show that reviews_count and average_review_length are the most dominant contributors to the model. This suggests that the reviewer's behavioral history is related to information richness and serves as a reliable indicator of helpfulness. Our study found that reviewer reputation can be linked to a review's helpfulness rating. A reviewer with a history of detailed and informative reviews is more likely to have higher helpfulness scores. This also suggests that readers rely on the reputation of an influential reviewer when determining the helpfulness of a review, especially when it is difficult to judge the quality of a product from the product description alone. The feature importance score of both models shows that the review length moderately influences the perceived helpfulness of a review. Review length can also indicate a helpful review, as longer reviews are more likely to contain beneficial content. Hence, it generally leads to more positive and helpful ratings from readers. The Feature importance of both models shows a low contribution by the verified purchase and rating of the review. Our study encodes the level of information regarding the product by matching the review content with the

product description, features, and categories. Similarity between review and product categories shows a moderate but consistent role in both gradient boosting and Random Forest classifiers, although Gradient Boosting does not rely on it. This semantic feature measures how well a review aligns with the core theme of the product. Other semantic-based features display very little contribution to both models, which may be because not all products have features and descriptions provided with the product metadata. The results of our study indicated that the informativeness of review content, along with the reviewer attributes, also contributes to the helpfulness rating of the review.

Since their AUC scores are identical, we can conclude that, for better precision, Gradient Boosting is suitable. However, for balanced precision and recall, Random Forest's use of multiple features makes it more desirable for generalizing over edge cases.

For the Fake review classification, the hyperparameter tuning process of 3 fold grid search allowed to search the best parameters for AdaBoost classifier out of a parameters grid comprising of Estimator criterion: ['gini', 'entropy'], estimator splitter": ["best", "random"], and n_estimators: [1, 2], the best parameters found are 'estimator__criterion': 'entropy', 'estimator__splitter': 'random' and 'n_estimators': 2. For Random Forest classifier the out of a parameter grid comprising of bootstrap: [True], max_depth: [80, 90, 100, 110], max_features: [2, 3], min_samples_leaf : [2, 3, 4], min_samples_split: [2, 5, 10], n_estimators: [60, 90, 120], the best parameters obtained for Random Forest are 'bootstrap'=True, 'max_depth': 80, 'max_features'= 3, 'min_samples_leaf'=4, 'min_samples_split ':10 and 'n_estimators': 120.

Table 11. Classification Results for Fake Reviews Classification

	AdaBoost	Random Forest
Accuracy	0.71	0.848
Precision	0.88	0.86
Recall	0.76	0.97
F1-Score	0.81	0.91
AUC Score	0.64	0.79

Note: Evaluation scores display that the Random Forest classifier outperforms the AdaBoost classifier in terms of accuracy, recall, F1-Score, and AUC Score. Evaluation metrics were calculated by applying the methodology defined in Equations 10 - 13.

The results of the classification of fake reviews are shown in Table 11 for both the trained classifiers. The Random Forest and AdaBoost models were able to achieve 85% and 71% accuracy in predicting the authenticity of a review, respectively. The confusion matrices for the AdaBoost classifier and the Random Forest classifier for fake review classification are given in Table 12 and Table 13, respectively.

The strength of AdaBoost lies in its higher precision, making it better at avoiding false positives; however, it also misses many fake reviews, indicated by its low recall. In contrast, Random Forest shows high recall but slightly lower precision (more false positives), which is still an acceptable tradeoff for filtering fake reviews Random Forest shows high recall but slightly lower precision (more false positives), which is still an acceptable tradeoff for filtering fake reviews.

Table 12. Confusion Matrix for AdaBoost Classifier for fake reviews classification

	Predicted 0	Predicted 1
Actual 0	67,422	66,791
Actual 1	158,074	506,569

Note: The above confusion Matrix indicates strong model performance with significantly high true positives but struggles in comparison to correctly identifying genuine reviews, misclassifying 66,791 as fake. Class 0 stands for honest reviews, and Class 1 represents fake reviews.

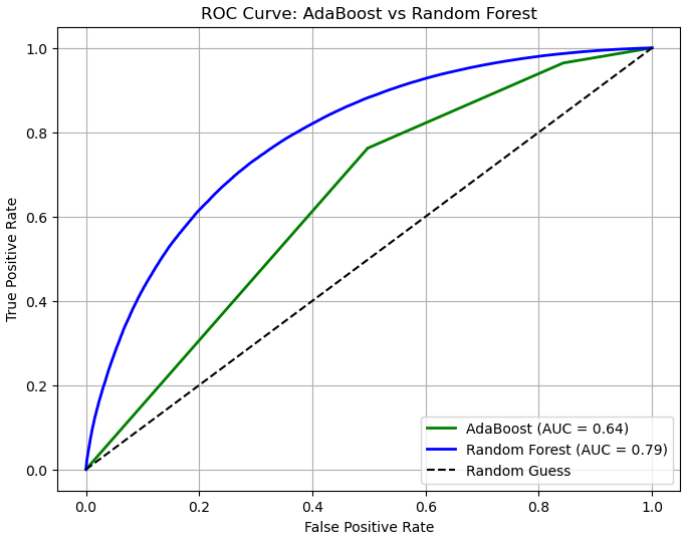
Table 13. Confusion Matrix for Random Forest Classifier for fake reviews classification

	Predicted 0	Predicted 1
Actual 0	28,599	105,614
Actual 1	14,845	649,798

Note: The above confusion Matrix indicates strong model performance with significantly high true positives but struggles in comparison; it also misclassifies many real reviews as fake, thereby indicating a tradeoff between precision and recall. The Class 0 stands for real reviews, and Class 1 represents fake reviews.

The confusion matrices in Table 12 and Table 13 reveal a significant difference in their manner of fake review classification for both models. AdaBoost achieves a large number of true positives but misclassifies many genuine reviews as fake. In contrast, Random Forest delivers a higher number of true positives but also misclassifies a larger number of genuine reviews as fake. However, AdaBoost has lower true positives out of all positive cases, indicating that it misses a substantial number of fake reviews. This specifies the effectiveness of Random Forest in filtering out fake reviews.

Figure 6. ROC curve comparison between AdaBoost Classifier(clf1) and Random Forest Classifier(clf2) for Fake reviews Detection



Note: The number of false positives is given along the x-axis, while the true positive rate is given along the y-axis. The diagonal dashed line denotes the random guess baseline with AUC 0.5. Random Forest classifier provides a better AUC of 0.79 than the AdaBoost Classifier with 0.64 AUC. AUC scores were calculated by applying the methodology defined in Equations (14-15).

The ROC curve comparison provided in Figure 6 provides additional insight into the discriminative ability of both classifiers. Both classifiers perform better than random guessing, implying they successfully note the meaningful patterns that discriminate between fake and genuine reviews. The higher AUC of Random Forest highlights the superiority of Random Forest in classifying correctly across different thresholds.

While AdaBoost only suggests a modest improvement over the random guess baseline. Hence, ROC analysis reinforces the earlier findings, confirming its effectiveness in filtering fake reviews effectively. Hence, Random Forest was chosen for fake review detection due to its superior AUC score of 0.79 compared to AdaBoost's 0.64.

The assessment of statistical significance of features for fake reviews classification for the AdaBoost classifier is provided in Table 14, and for the Random Forest model is provided in Table 15.

Table 14. Permutation Importance with one-sample t-test scores for AdaBoost Classifier fake reviews classification

Feature	Mean Importance	CI Lower	CI Upper	p-adj
Sentiment	0.252360***	0.251909	0.252812	p < 0.001
Neg_Count	0.122691***	0.122301	0.123080	p < 0.001
Authenticity	0.105790***	0.105444	0.106136	p < 0.001
Word_Count	0.100550***	0.100269	0.100830	p < 0.001
Unique_words	0.099702***	0.099312	0.100091	p < 0.001
Subjectivity	0.095990***	0.095641	0.096339	p < 0.001
Pro_Count	0.059122***	0.058800	0.059444	p < 0.001
Aux_Count	0.048152***	0.047858	0.048446	p < 0.001
Verb_Count	0.045621***	0.045382	0.045861	p < 0.001
AT	0.045206***	0.044942	0.045471	p < 0.001
Adv_Count	0.043618***	0.043345	0.043891	p < 0.001
Pre_Count	0.042209***	0.041919	0.042498	p < 0.001
Noun_Count	0.040078***	0.039828	0.040328	p < 0.001
Adj_Count	0.039802***	0.039561	0.040043	p < 0.001
Art_Count	0.036670***	0.036439	0.036900	p < 0.001
Con_Count	0.034579***	0.034346	0.034813	p < 0.001

Note: *** indicates statistical significance at $p < 0.001$ after Bonferroni correction. All features are statistically significant with p-adj < 0.05. The strongest predictors are Sentiment, Neg_Count, Authenticity, and Word_Count, while part-of-speech level counts contributed marginally. It implies that semantic and stylistic features dominate over basic lexical counts for fake review classification. Permutation Importance metrics were calculated by applying the methodology defined in Equation (19).

For the AdaBoost classifier, as shown in Table 14, sentiment permutation mean is the most influential feature, which suggests that the polarity of a review significantly determines the credibility of the review. Similarly, the number of negative words in a review signals either dissatisfaction or deceptive review behavior. Authenticity is another highly influential feature, indicating that stylometric cues of sincerity matter substantially. Another implication from the permutation importance is that length and diversity matter in differentiating between valid and invalid reviews. The stylistic feature of subjectivity is also another influential predictor. The structural features, such as pronoun use, verb usage, etc., do not dominate individually, but collectively they can help differentiate credible reviews from fraudulent reviews. Prior approaches mostly focused only on numeric features, sentiment, or review attributes, while our work leveraged linguistic and sentiment features to provide a more comprehensive representation of reviews.

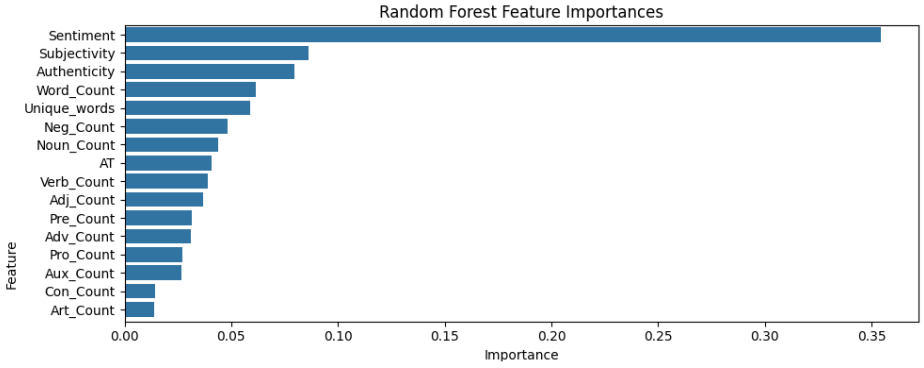
Table 15. Permutation Importance with one-sample t-test scores for Random-Forest Classifier fake reviews classification

Feature	Mean Importance	CI Lower	CI Upper	p-adj
Sentiment	0.077388***	0.077114	0.077663	p < 0.001
Authenticity	0.018304***	0.018175	0.018434	p < 0.001
Word Count	0.016818***	0.016656	0.016980	p < 0.001
Neg Count	0.015348***	0.015230	0.015467	p < 0.001
Subjectivity	0.015181***	0.015049	0.015314	p < 0.001
Unique Words	0.014120***	0.014022	0.014219	p < 0.001
Verb Count	0.010694***	0.010595	0.010792	p < 0.001
AT	0.009743***	0.009652	0.009834	p < 0.001
Adv Count	0.009726***	0.009626	0.009825	p < 0.001
Aux Count	0.009137***	0.009064	0.009211	p < 0.001
Noun Count	0.008024***	0.007926	0.008121	p < 0.001
Adj Count	0.007692***	0.007617	0.007767	p < 0.001
Pro Count	0.007608***	0.007516	0.007699	p < 0.001
Pre Count	0.007452***	0.007381	0.007524	p < 0.001
Con Count	0.005835***	0.005760	0.005910	p < 0.001
Art Count	0.005014***	0.004933	0.005095	p < 0.001

Note: *** indicates statistical significance at p < 0.001 after Bonferroni correction. All features are statistically significant with p-adj < 0.05. The strongest predictors are Sentiment, Neg_Count, Authenticity, and Word_Count, while part-of-speech level counts contributed marginally. It implies that semantic and stylistic features dominate over basic lexical counts for fake review classification. Permutation Importance scores were calculated by applying the methodology defined in Equation (19).

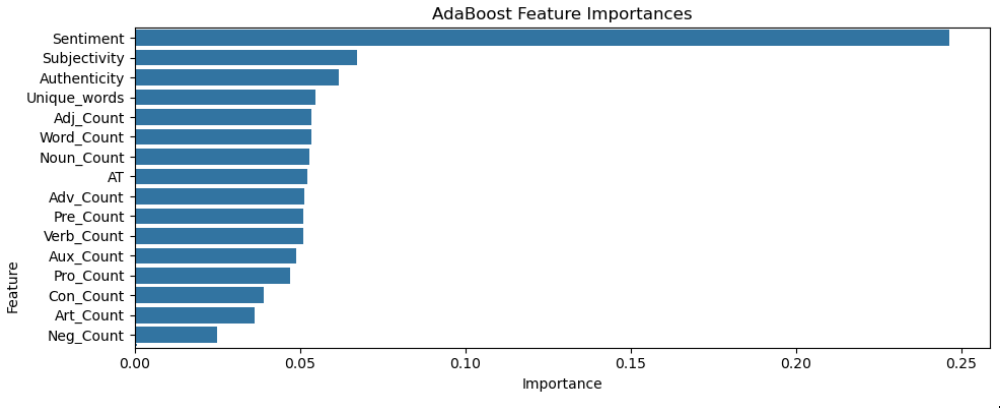
For the Random Forest model, as shown in Table 15, sentiment stands as the most impactful feature with the mean importance value of 0.0774, confirming that polarity remains a significant factor for fake review classification. The subsequent predictors are Authenticity, Word Count, Negative Word Count, Subjectivity, and Unique Words, which establish stylometric sincerity. Hence, review length, presence of negative expressions, perceived subjectivity, and lexical richness collectively impact the credibility. Compared to the Ada-Boost model, the ordering shows greater balance between semantic-stylistic features and basic lexical counts. Features like Verb Count, AT, Adverb Count, Auxiliary Count, Noun Count, and Adjective Count, although individually non-dominant, still impact and play a pivotal role in differentiating patterns between fake and genuine reviews. Similarly, features like Pronoun Count, Preposition Count, Conjunction Count, and Article Count, with the lowest scores, have a weaker but non-negligible role.

Figure 7. Feature Importances for Random Forest Classifier for fake reviews classification



Note: The Figure shows feature importance for the Random Forest model on the x-axis and feature name on the y-axis. Feature importance metrics were calculated by applying the methodology defined in Equation (16).

Figure 8. Feature Importances for AdaBoost Classifier for fake reviews classification



Note: The Figure shows feature importance for the AdaBoost model on the x-axis and feature name on the y-axis. Feature importance metrics were calculated by applying the methodology defined in Equation (18).

A bar plot in Figures 7 and 8 visualizes feature importance in a Random Forest classifier and an AdaBoost classifier, respectively, showing the contribution of each feature in making predictions and achieving the current scores. Overall, the most dominant feature for both classifiers is the sentiment score of the review. The features related to review content structure, such as word count, unique word count, authenticity, and subjectivity, also positively influence both classifiers. Again, Random Forest shows a broader distribution of contributions from several features. This demonstrates that Random Forest has better flexibility in utilizing semantic and linguistic aspects.

The strength of our work is evident in Table 16, which presents a comparative analysis of our findings with the state-of-the-art research.

Table 16. Comparison of our work with previous research

	Targeted Problem	Features	Dataset	Classifier	Accuracy
Our Work	Helpful review detection	Review text, Review Sentiment, Reviewer, and Product attributes	Publicly available Amazon Dataset 120,000 reviews, Review annotated as helpful if helpful vote count > 10	Random Forrest	94%
				Gradient Boosting	94%
	Fake Reviews Detection	Word count, sentiment, linguistic features	Amazon dataset "Toys and Games" for fake reviews	AdaBoost	71
				Random Forest	85%
(Periasamy et al., 2024)	Fake Reviews Detection	Sentiment	Manually labelled Custom Datasets	BERT ensemble (Decision Trees, SVM, KNN)	80%

(Choi et al., 2023)	Usefulness detection	Word count, sentiment score, length	Naver shopping mall platform No.03 "Teeth-Whitening" merchandise, 4,000 reviews were crawled.	Support Vector Classifier	85%
	Fake Reviews Detection	Content quality based on pros and cons, multimedia metadata		KNN for Fake Review Detection	N/A
(Bilal & Almazroi, 2023)	Helpful Review detection	Fine-Tuned BERT	Yelp Dataset 10,000 reviews, Review annotated as helpful if helpful vote count >= 4	BERT Classifier	70%
(Zhou & Yang, 2019)	Helpful Review detection	Rating, Length, Sentiment, complexity, reputation, ranking, age	Scrapped Amazon Dataset, 30,338 reviews, Review annotated as helpful if helpful vote count > 0	Random Forest	80%

Note: Comparison with previous work is provided in terms of the targeted problem, features used, dataset size, classification technique, and the achieved score.

The research remains limited to only one aspect of review analysis, either in terms of the usefulness of reviews or the detection of fake reviews. Only a few scholars (Choi et al., 2023) have provided similar research, albeit with limited data samples and no publicly available dataset. Our work displays significant accuracy for both usefulness and fake review classifications. The study enhanced the classification performance of current approaches, even when a more extensive dataset was used compared to previous ones.

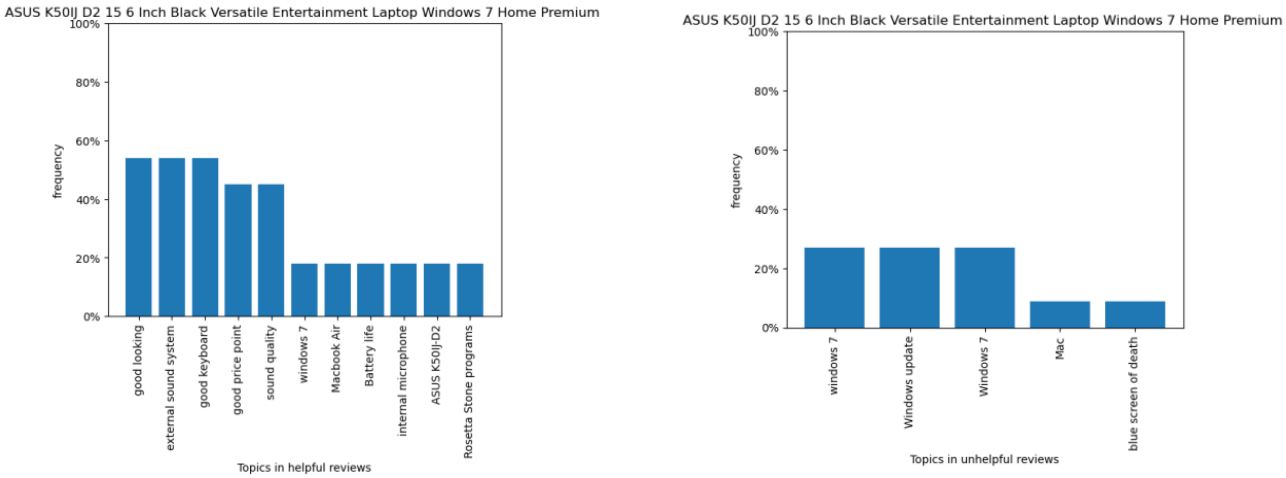
5.1 Data Analysis Topic Identification

Helpful reviews are crucial for identifying topics and understanding customer needs and preferences. By withdrawing topics from reviews that are considered helpful, industries can increase understanding of the features of their products or facilitate connections that positively connect with customers. This material enables businesses to tailor their services to meet customer expectations, ultimately enhancing customer satisfaction and loyalty.

Dividing the topic identification into negative and positive sentiment reviews yields a more detailed understanding of customer responses. By recognizing topics exactly stated in these reviews, industries can identify areas that need improvement or discover recurring issues to prevent customer dissatisfaction. Conversely, recognizing topics in positive sentiment reviews helps industries identify their durability and areas of success.

For the extraction of meaningful topics from customer reviews, we employed Google's Large Language Model, Gemini. An input of review text was provided, along with a prompt to generate topics from the reviews as output. These input-output pairs of at least 43 reviews were selected by the investigator and then stored as in-context guidance, along with prompts for the model. These pairs are publicly available at an online repository (Aslam, 2025). For example, a review “The battery lasts long, and the device runs smoothly” was paired with output topics “Battery, performance”. These pairs were used to construct a consistent prompt structure and were passed to the Gemini API using gemini-1.0-pro for topic extraction for all reviews of a single product. A Python script iterated over each review of the selected product from over 10,000 reviews, took prompt arrays as context, and returned topic predictions for each review by the Gemini Model.

Figure 9. Topic analysis of helpful and unhelpful reviews

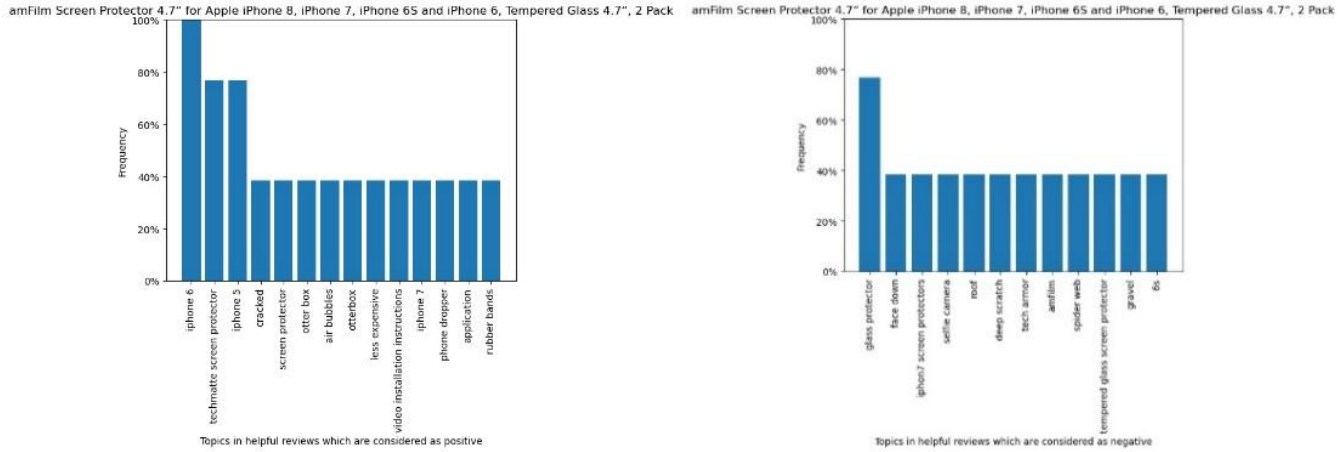


(a) Most discussed topics in the helpful reviews **(b)** Most discussed topics in the unhelpful reviews

Note: The frequency of a topic in percentage is given along the y-axis, and the topics are provided along the x-axis. A comparison of the most frequently discussed topics among reviews classified as helpful in Figure 9 (a) with the most discussed topics among the unhelpful reviews in Figure 9 (b) can be seen.

Figures 9 (a) and (b) display the most discussed topics in a product, helpful and unhelpful reviews, demonstrating that helpful reviews are more relevant to the product than unhelpful reviews. We have selected BERT for sentiment classification, which is supported by research (Elmitwalli & Mehegan, 2024). We separated reviews into those with negative and positive sentiment for a more nuanced understanding of customer feedback. Then, topics were extracted from Negative sentiment reviews to highlight areas where products or services may be lacking or where improvements are needed. The exact process was also performed for the positive reviews.

Figure 10. Topic analysis of positive and negative helpful reviews



(a) Most discussed topics in the positive and helpful reviews

(b) Most discussed topics in the negative and helpful reviews

Note: The frequency of a topic in percentage is given along the y-axis, and the topics are provided along the x-axis. A comparison of the most frequently discussed topics among positive reviews classified as helpful in Figure 10 (a) with the most discussed topics among negative reviews classified as helpful in Figure 9 (b) can be seen.

Figures 10 (a) and (b) demonstrate the sentiment-based topic analysis of a product in helpful reviews. Such analysis can help businesses pinpoint areas for improvement or address recurring issues to prevent customer dissatisfaction by identifying topics mentioned in these reviews.

For example, Figure 9 presents a review summary for the ASUS K50IJ laptop, including both helpful and unhelpful reviews. The helpful reviews highlight main aspects such as sound quality, keyboard design, and battery life. In contrast, the unhelpful reviews cover generic topics, including Windows 7 and operating system-related terms, as the most frequent subjects. By retaining only helpful reviews, researchers and businesses can obtain a clearer picture of customer priorities. Moreover, Figure 10 demonstrates that the value of helpful reviews can be further understood when their sentiment polarity is also considered as context. For example, the positive helpful reviews of screen protectors dominantly discuss installation instructions, the names of multiple iPhone models, and packaging. In contrast, the negative helpful reviews discuss adhesive failure, durability issues, and selfie camera problems.

Finally, the study demonstrated how helpful content can be analyzed to understand customer needs. A frequency-based topic graph was created to reveal the most significant topic with frequency among the helpful reviews. Insight from the consumers' reactions offers a wealth of information for tailored marketing and suggestions. This research thoroughly investigated what was needed to determine whether relevant features from consumer comments can be extracted to support the quantitative application of classical preference identification and product improvement models, despite the fact that these models already exist and provide a theoretical foundation for this research. This research combined these approaches based on improved user-generated content feature mining techniques to suggest products to users more effectively, which is a promising avenue for future study. Linguistic analysis of review content

may further strengthen the study. Also, another possible avenue for further research in this area is how to present reviews differently depending on user preferences and personal data.

An essential impact of the findings is that online businesses should invest more time and effort in managing reviews, numerical data, sentiment, reviewers, and product elements on their web pages to gain more convenient insights into customer needs and responses. The practical implication of this study is that online retailers and businesses should formulate rules to avoid leaving the categories or features section empty regarding a product. Additionally, reviews on most platforms are not displayed with the required reputation score of the reviewer. A positive review by an influential reviewer can be listed at the top of the reviews to impact product revenue (Chua & Banerjee, 2015). Additionally, a guiding note can be provided to the writer while writing a review to help them craft more impactful reviews, leading to a deeper understanding of the consumer's opinion.

In summary, we address some of the core challenges of automated content moderation through our proposed feature extraction technique, which significantly contributes to the advancement of both helpful review and fake review detection, as well as their practical implications. Future work can focus on incorporating images and videos to enhance the understanding of the review.

5.2 Managerial Implications

Helpful reviews are crucial for customers on e-commerce platforms when making buying decisions. Hence, such helpful reviews can aid in identifying the content that buyers find most helpful and recognizing customers' needs. Since most reviews on the platform remain unrated, it is valuable research to build an automatic helpfulness assessment system capable of recognizing the type of content in reviews that customers find helpful. For this reason, this research provides insight into understanding the multi-layered aspects of helpful reviews. To this end, this study has multifaceted managerial implications. First, this analysis of helpful reviews provides strategic information for brands using e-commerce platforms to gain an in-depth understanding of customer needs, inclinations, and real-time recurring issues. Therefore, the analysis provides a comprehensive model to help brands develop a monitoring mechanism for a vibrant feedback loop. This proactiveness can benefit brands using e-commerce by tailoring their offerings and enhancing customer experiences, leading to notable customer fulfillment and loyalty. Particularly in the digital era, feedback is critical for brands operating online. Brands reflect their commitment through timely addressing of issues, and monitoring can swiftly prevent minor issues from escalating and respond by tailoring advertising and marketing communications.

Secondly, helpful reviews often highlight gaps in products and services, identifying innovative ideas for brands that are constantly seeking new ideas. Through timely monitoring and evaluation of such helpful reviews and feedback, brands can concentrate on quality or developments that precisely address customers' needs, giving them a competitive advantage. Thirdly, for strategic communication management of a brand in the digital era, helpful reviews can offer valuable insights into how customers perceive brand communication and promotional efforts. This enables brands to refine their marketing communication strategies by focusing on tailored communication campaigns on e-commerce platforms that resonate

primarily with their target customers. In particular, an analysis of the sentiments that raise concerns about the product can help protect its reputation on e-commerce platforms. Lastly, insights gained from sentiment analysis regarding the emotional tone of customers' reviews can help brands segment their customer base more precisely, steering them toward additional pursued and customized brand messaging and marketing strategies that increase customer engagement. This study suggests that integrating informed (e.g., review analysis-driven) strategies enables brands to remain agile, fostering progress and resilience in a competitive e-commerce arena.

6. Conclusions

This study aims to investigate the potential of ML for predicting the validity and helpfulness of online reviews, as they are a critical factor in shaping managerial strategies as well as creating customer opinions and purchase decisions in e-commerce platforms. The increase of fake and unhelpful reviews on e-commerce platforms causes significant information overload, leading to an impact on the decision-making process. Earlier studies have considered review helpfulness and fake review detection in isolation, often relying on limited datasets and features. An integrated and comprehensive framework can jointly address authenticity and help in detecting review evaluation.

The motivation behind our study is the research gap to develop and validate a two-stage, data-driven framework that filters out unhelpful and fake reviews using textual, semantic, reviewer-level, and product-level features. Specifically, the framework integrates sentiment polarity, word diversity, parts-of-speech analysis for verifying the credibility of reviews, while utilizing review and product metadata, and semantic similarity using SBERT embeddings for examining the informativeness of a review. In addition, this study implements a sentiment-based topic extraction model leveraging Gemini to derive actionable insights from validated reviews.

The empirical findings demonstrate strong predictive performance across six Amazon product categories for the helpfulness classification. The Random Forest classifier achieved 94% accuracy, precision, and F1-score, with 93% recall and an AUC score of 98%, while Gradient Boosting also attained comparable results for helpfulness classification. For fake review detection, the Random Forest classifier attained 85% accuracy, 86% precision, 97% recall, and 91% F1-score with an AUC score of 79%. These results validate the robustness and generalizability of the proposed framework. Prominently, the findings indicate that combining semantic similarity with product metadata, reviewer credibility metrics, and linguistic features significantly enhances the reliability of review quality assessment.

These findings lead to the inference that semantic alignment between review content and product metadata supplies contextual relevance and serves as a strong indicator of review usefulness, while linguistic features influence perceived credibility. Also, reviewer-level behavioral indicators substantially improve helpfulness detection performance, highlighting the importance of integrating textual and non-textual features in predictive systems.

From an academic perspective, this study contributed to the field of decision sciences and to the literature of information systems with its novel dual-task data-driven framework that evaluates review authenticity and helpfulness to obtain reliable reviews. Unlike previous literature that examined these problems in isolation, this research integrates them into a unified architecture and validates it using a large publicly available dataset. The findings provide empirical support and methodological foundation for future theory development in online consumer behavior and content credibility assessment.

The findings of our study have several practical implications. From a managerial perspective, it is automating the identification and filtering of fraudulent and unhelpful reviews, which reduces informational noise in e-commerce platforms and leads to enhanced trust in user-generated content. Furthermore, our sentiment-based topic extraction applied to validated reviews enables managers to develop an understanding of customer perceptions regarding product attributes, thus assisting in informed marketing policies and decision-making processes.

Despite its contributions, this study has certain limitations. First, the analysis of this study is restricted to English-language reviews from the Amazon platform, limiting the generalizability across different cultural and linguistic backgrounds. Second, the framework focuses solely on the textual content of reviews and metadata features, excluding multimodal elements from reviews, such as images, emojis, and videos, that may provide additional signals of authenticity and helpfulness.

Therefore, this study provides several recommendations for future studies. Future studies may extend this framework to multilingual datasets to strengthen generalization. Incorporating multimodal review content and buyer behavioral profiles may be investigated further to identify their role in enhancing detection accuracy. Moreover, real-time deployment through an Application Programming Interface (API) in a dynamic and large-scale environment can be utilized to improve the reliability and performance of recommendation engines in future studies.

References

Abd, M. J., & Hussein, M. H. (2024). Fake reviews detection in e-commerce using machine learning techniques: a comparative survey. *BIO Web of Conferences*, *97*. https://doi.org/10.1051/bioconf/20249700099

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102–4107. https://doi.org/10.3923/jeasci.2017.4102.4107

Aslam, F. (2025). *TopicExamples*. GitHub. https://github.com/FarrahAslam110/TopicExamples

Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial Intelligence Review*, *56*(9), 10345–10425. https://doi.org/10.1007/s10462-023-10419-1

Bilal, M., & Almazroi, A. A. (2023). Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews. *Electronic Commerce Research*, *23*(4), 2737–2757. https://doi.org/10.1007/s10660-022-09560-w

Bilal, M., Marjani, M., Hashem, I. A. T., Gani, A., Liaqat, M., & Ko, K. (2019). Profiling and predicting the cumulative helpfulness (Quality) of crowd-sourced reviews. *Information (Switzerland)*, *10*(10). https://doi.org/10.3390/info10100295

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Camacho-Otero, J., Boks, C., & Pettersen, I. N. (2019). User acceptance and adoption of circular offerings in the fashion sector: Insights from user-generated online reviews. *Journal of Cleaner Production*, *231*(231), 928–939. https://doi.org/10.1016/j.jclepro.2019.05.162

Campos, P., Pinto, E., & Torres, A. (2025). Rating and perceived helpfulness in a bipartite network of online product reviews. *Electronic Commerce Research*, *25*(3), 1607–1639. https://doi.org/10.1007/s10660-023-09725-1

Changchit, C., & Klaus, T. (2020). Determinants and Impact of Online Reviews on Product Satisfaction. *Journal of Internet Commerce*, *19*(1), 82–102. https://doi.org/10.1080/15332861.2019.1672135

Chatterjee, S. (2025). Effect of construal level on the drivers of online-review-helpfulness. *Electronic Commerce Research*, *25*(2), 1115–1143. https://doi.org/10.1007/s10660-023-09716-2

Chen, C., Zhou, J., Qiu, M., Li, X., Bao, F. S., Yang, Y., & Huang, J. (2019). Multi-domain gated CNN for review helpfulness prediction. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2630–2636. https://doi.org/10.1145/3308558.3313587

Chen, L., Chen, G., & Wang, F. (2015). Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*, *25*(2), 99–154. https://doi.org/10.1007/s11257-015-9155-5

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cheng, Y., Hui, Y., Liu, S., & Wong, W. K. (2022). Could significant regression be treated as insignificant: An anomaly in statistics? *Communications in Statistics Case Studies Data Analysis and Applications*, *8*(1), 133–151. https://doi.org/10.1080/23737484.2021.1986171

Cheng, Y., Hui, Y., McAleer, M., & Wong, W. K. (2021). Spurious Relationships for Nearly Non-

Stationary Series. *Journal of Risk and Financial Management*, *14*(8). https://doi.org/10.3390/jrfm14080366

Choi, W., Nam, K., Park, M., Yang, S., Hwang, S., & Oh, H. (2023). Fake review identification and utility evaluation model using machine learning. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.1064371

Chua, A. Y. K., & Banerjee, S. (2015). Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth. *Journal of the Association for Information Science and Technology*, *66*(2), 354–362. https://doi.org/10.1002/asi.23180

Daniels, M. (2022). *(2022, 10). Why the FTC is trying to crack down on fake reviews on e-commerce sites. (modernretail)*. https://www.modernretail.co/operations/why-the-ftc-is-trying-to-crack-down-on-fake-reviews-on-e-commerce-sites/

Deldjoo, Y., Nazary, F., Ramisa, A., Mcauley, J., Pellegrini, G., Bellogin, A., & Noia, T. D. (2023). A review of modern fashion recommender systems. *ACM Computing Surveys*, *56*(4), 1-37. https://doi.org/10.1145/3624733

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427–431.

Duma, R. A., Niu, Z., Nyamawe, A. S., Tchaye-Kondi, J., Jingili, N., Yusuf, A. A., & Deve, A. F. (2024). Fake review detection techniques, issues, and future research directions: a literature review. *Knowledge and Information Systems*, *66*(9), 5071–5112. https://doi.org/10.1007/s10115-024-02118-2

Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1). https://doi.org/10.1186/s40649-020-00080-x

Elmitwalli, S., & Mehegan, J. (2024). Sentiment analysis of COP9-related tweets: a comparative study of pre-trained models and traditional techniques. *Frontiers in Big Data*, *7*. https://doi.org/10.3389/fdata.2024.1357926

Enamul Haque, M., Tozal, M. E., & Islam, A. (2018). Helpfulness prediction of online product reviews. *Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018*, 1–4. https://doi.org/10.1145/3209280.3229105

Fan, M., Feng, Y., Sun, M., Li, P., Wang, H., & Wang, J. (2018). Multi-task neural learning architecture for end-to-end identification of helpful reviews. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, 343–350. https://doi.org/10.1109/ASONAM.2018.8508623

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Hannan, S. A., Ahmed, S. J., Naveed, Q., & Thakur, R. A. (2012). Data Mining and Natural Language Processing Methods for Extracting Opinions from Customer Reviews. *International Journal of Computational Intelligence and Information Security*, *3*(6), 52–58.

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2024). *Bridging Language and Items for Retrieval and Recommendation*. http://arxiv.org/abs/2403.03952

Hussain, N., Turab Mirza, H., Hussain, I., Iqbal, F., & Memon, I. (2020). Spam Review Detection Using the Linguistic and Spammer Behavioral Methods. *IEEE Access*, *8*, 53801–53816. https://doi.org/10.1109/ACCESS.2020.2979226

INC42. (2024). *Ecommerce Platforms Bat For Crackdown On Fake Reviews. (StartupNews.fyi)*. https://startupnews.fyi/2024/05/16/ecommerce-platforms-bat-for-crackdown-on-fake-reviews/

Işik, M., & Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(3), 1405–1421. https://doi.org/10.3906/elk-1907-46

Jyoti, S. D., & Singh, K. (2015). Comparison of various similarity measure techniques for generating recommendations for E-commerce sites and social websites. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, *11*(2), 219-221. http://www.iasir.net

Koroteev, M. V. (2021). BERT: A Review of Applications in Natural Language Processing and Understanding. *ArXiv(2103.11943)*. http://arxiv.org/abs/2103.11943

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

Kühl, N., Mühlthaler, M., & Goutier, M. (2020). Supporting customer-oriented marketing with artificial intelligence: automatically quantifying customer needs from social media. *Electronic Markets*, *30*(2), 351–367. https://doi.org/10.1007/s12525-019-00351-0

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-y

Li, Q., Park, J., & Kim, J. (2024). Impact of information consistency in online reviews on consumer behavior in the e-commerce industry: a text mining approach. *Data Technologies and Applications*, *58*(1), 132–149. https://doi.org/10.1108/DTA-08-2022-0342

Li, S., Liu, F., Zhang, Y., Zhu, B., Zhu, H., & Yu, Z. (2022). Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review. *Mathematics*, *10*(19). https://doi.org/10.3390/math10193554

Lin, X. (2020). Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing. *ACM International Conference Proceeding Series*, 32–36. https://doi.org/10.1145/3436286.3436293

Mumuni, A. G., O'Reilly, K., MacMillan, A., Cowley, S., & Kelley, B. (2020). Online Product Review Impact: The Relative Effects of Review Credibility and Review Relevance. *Journal of Internet Commerce*, *19*(2), 153–191. https://doi.org/10.1080/15332861.2019.1700740

Nguyen, N. M., Huong Giang, H., Vu, N. T. M., & Ta, S. A. (2025). How do online reviews moderate effects of country image on product image and purchase intention: cases of Korean and US products in Vietnam. *Asia-Pacific Journal of Business Administration*, *17*(2), 337–358. https://doi.org/10.1108/APJBA-07-2023-0346

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Park, K., Hong, J. S., & Kim, W. (2020). A Methodology Combining Cosine Similarity with Classifier for Text Classification. *Applied Artificial Intelligence*, *34*(5), 396–411. https://doi.org/10.1080/08839514.2020.1723868

Paul, H., & Nikolaev, A. (2021). Fake review detection on online E-commerce platforms: a systematic literature review. *Data Mining and Knowledge Discovery*, *35*(5), 1830–1881. https://doi.org/10.1007/s10618-021-00772-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Periasamy, M., Mahadevan, R., Raman, R. C., & Jessiman, J. (2024). Finding fake reviews in e-commerce platforms by using hybrid algorithms. *arXiv preprint arXiv:2404.06339*.

Qiu, K., & Zhang, L. (2024). How online reviews affect purchase intention: A meta-analysis across contextual and cultural factors. *Data and Information Management*, *8*(2), 100058. https://doi.org/10.1016/j.dim.2023.100058

Racherla, P., & Friske, W. (2012). Perceived "usefulness" of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications*, *11*(6), 548–559. https://doi.org/10.1016/j.elerap.2012.06.003

Ranfagni, S., & Rosati, M. (2023). Triangulating online brand reputation, brand image, and brand identity: An interdisciplinary research approach to design the pathways of online branding strategies in luxury hospitality. In *Online Reputation Management in Destination and Hospitality: What We Know, What We Need To Know*. Emerald Publishing Limited. https://doi.org/10.1108/978-1-80382-375-120231012

Ren, G., & Hong, T. (2019). Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. *Information Processing and Management*, *56*(4), 1425–1438. https://doi.org/10.1016/j.ipm.2018.04.003

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, *64*, 102771. https://doi.org/10.1016/j.jretconser.2021.102771

Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A Review of Applications in Sentiment Analysis. *HighTech and Innovation Journal*, *4*(2), 453–462. https://doi.org/10.28991/HIJ-2023-04-02-015

Shahmirzadi, O., Lugowski, A., & Younge, K. (2019). Text similarity in vector space models: A

comparative study. *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, 659–666. https://doi.org/10.1109/ICMLA.2019.00120

Shirkhani, S., Mokayed, H., Saini, R., & Chai, H. Y. (2023). Study of AI-Driven Fashion Recommender Systems. *SN Computer Science*, *4*(5), 514. https://doi.org/10.1007/s42979-023-01932-9

Singh, A., & Garg, S. K. (2024). Comparative Study of Different Document Similarity Measures and Models. In *Lecture Notes in Networks and Systems* (Vol. 894). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9562-2_61

Singh, H., Chakrabarti, S., & Utkarsh. (2023). How do gratifications to read reviews and perceived reviewers' credibility impact behavioural intentions in fashion e-commerce? A mediating-moderating perspective. *Computers in Human Behavior*, *143*(4), 107677. https://doi.org/10.1016/j.chb.2023.107677

Singh, U., Saraswat, A., Azad, H. K., Abhishek, K., & Shitharth, S. (2022). Towards improving e-commerce customer review analysis for sentiment detection. *Scientific Reports*, *12*(1), 21983. https://doi.org/10.1038/s41598-022-26432-3

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science* (Vol. 11856, pp. 194–206). Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Sun, X., Han, M., & Feng, J. (2019). Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products. *Decision Support Systems*, *124*, 113099. https://doi.org/10.1016/j.dss.2019.113099

Sung, E., Chung, W. Y., & Lee, D. (2023). Factors that affect consumer trust in product quality: a focus on online reviews and shopping platforms. *Humanities and Social Sciences Communications*, *10*(1), 1–10. https://doi.org/10.1057/s41599-023-02277-7

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*, 5999–6009.

Wong, W. K., Cheng, Y., & Yue, M. (2024). Could Regression of Stationary Series Be Spurious? *Asia-Pacific Journal of Operational Research*. https://doi.org/10.1142/S0217595924400177

Wong, W. K., & Pham, M. T. (2022a). Could the test from the standard regression model could make significant regression with autoregressive noise become insignificant—a note. *The International Journal of Finance*, *35*, 19–39.

Wong, W. K., & Pham, M. T. (2022b). Could the test from the standard regression model could make significant regression with autoregressive noise become insignificant. *The International Journal of Finance*, *34*, 1–18.

Wong, W. K., & Pham, M. T. (2023a). Could the test from the standard regression model could make significant regression with autoregressive Y_t and X_t become insignificant – a note. *The International Journal of Finance*, *35*, 20–41.

Wong, W. K., & Pham, M. T. (2023b). Could the test from the standard regression model could make significant regression with autoregressive Yt and Xt become insignificant. *The International Journal of Finance*, *35*, 1–19.

Wong, W. K., & Pham, M. T. (2025). How to model a simple stationary series with a non-stationary series. *The International Journal of Finance*, *37*, 1–19.

Wong, W. K., & Yue, M. (2024a). Could Regressing a Stationary Series on a Non-Stationary Series Obtain Meaningful Outcomes?- a remedy. *The International Journal of Finance*, *36*, 1–20. https://doi.org/10.1142/S2010495224500118

Wong, W. K., & Yue, M. (2024b). Could Regressing a Stationary Series on a Non-Stationary Series Obtain Meaningful Outcomes? *Annals of Financial Economics*, *19*(3), 2450011. https://doi.org/10.1142/S2010495224500118

Xu, S., Cuan, H., Yin, Z., & Yin, C. (2024). A Hybridized Approach for Enhanced Fake Review Detection. *IEEE Transactions on Computational Social Systems*, *11*(6), 7448–7466. https://doi.org/10.1109/TCSS.2024.3411635

Yin, D., Mitra, S., & Zhang, H. (2016). When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. *Information Systems Research*, *27*(1), 131–144. https://doi.org/10.1287/isre.2015.0617

Yoo, S. Y., & Jeong, O. R. (2020). Automating the expansion of a knowledge graph. *Expert Systems with Applications*, *141*, 112965. https://doi.org/10.1016/j.eswa.2019.112965

Zheng, L. (2021). The classification of online consumer reviews: A systematic literature review and integrative framework. *Journal of Business Research*, *135*, 226–251. https://doi.org/10.1016/j.jbusres.2021.06.038

Zhou, Y., & Yang, S. (2019). Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews. *IEEE Access*, *7*, 27769–27780. https://doi.org/10.1109/ACCESS.2019.2901472