A Novel Approach Based on Information Relevance Perspective and ANN for Predicting the Helpfulness of Online Reviews

Nur Syadhila Bt Che Lah[1], Khursiah Zainal-Mokhtar[2]
Department of Computer and Information Science, Universiti Teknologi PETRONAS, Perak, Malaysia[1]
Research Innovation Centre (RIC), Universiti Teknologi PETRONAS, Perak, Malaysia[2]

***Abstract*—This study presents a novel approach to predicting the helpfulness of online reviews using Artificial Neural Networks (ANNs) focused on information relevance. As online reviews significantly influence consumer decision-making, it is critical to understand and identify reviews that provide the most value. This research identifies four key textual features namely content novelty, content specificity, content readability, and content reliability, that contribute to perceived helpfulness and incorporates them as primary inputs for the ANN model. Datasets of Amazon reviews are analyzed, and various preprocessing steps are employed to ensure data quality. Reviews are classified as helpful or unhelpful based on helpful vote thresholds, with experiments conducted across multiple helpful vote thresholds to determine the optimal threshold value. Performance was evaluated using accuracy, precision, recall, and F1 scores, with the best-performing classifier achieving 74.34% accuracy at a helpful vote threshold of 12 votes. These results highlight the potential of information relevance-based criteria to enhance the accuracy of online review helpfulness prediction models.**

***Keywords*—Review helpfulness; online reviews; information relevance; review novelty; review readability; review specificity; Artificial Neural Networks**

I. INTRODUCTION

The emergence of internet has revolutionized life, bringing significant changes essential to people's daily life. In the past, tasks such as making purchases had to be done face-to-face. In this traditional market settings, consumers acquire new information via advertisements, brochures and word-of-mouth (WOM) about various products and services before making purchases. In the context of online reviews, consumer often reads multiple reviews before making purchasing decision. Online reviews have impacted not only consumers but also business, platforms and reviewers. Reviews can help consumers reduce uncertainty [1] regarding the quality of a product or service by offering firsthand insights and experiences from other users. For businesses, positive reviews can lead to increased sales as they build trust and credibility [2]. Platforms benefit from the continuous creation of online reviews and building consumer trust, which attracts more consumers. In turn, reviewers gain recognition from their peers and receive gifts and special promotion on platforms. Such incentives encourage them to continue providing useful reviews [3].

As the number of reviews rapidly increasing, platforms face the challenge of managing this large amount of information to ensure that consumers can easily access to the most relevant and helpful review. To address the issue, platforms have introduced feedback mechanisms that allow consumers to vote for posted reviews that they considered helpful. Readers are more likely to trust the statements if it has been marked as helpful by other consumers [4]. However, helpful vote is a manual process and could result in helpful reviews being ignored. Moreover, it is unclear to potential customers whether previous customers marked a review as "helpful" before using a product or service or after having used it. The criteria for what constitute a "helpful" review is not strictly defined and thus can be difficult to assess.

Many studies have utilized various theories such as Elaboration Likelihood Model (ELM) and Information Adoption Model (IAM) to capture numerous factors that affect review helpfulness [5-9]. In the context of IAM, it suggests that information helpfulness is influenced by argument quality and source credibility. Recent research has revealed inconsistencies on how to properly judge the quality of arguments in reviews. Some experts believe that certain parts of a review are difficult to measure objectively and can be changed depending on the situation [10]. Also, many traditional ways of measuring argument quality might not work well for evaluating personal, subjective aspects of the review content [11]. As a result, relevance has been suggested as an important factor, with its importance depending on the specific decision a reader is trying to make [12]. Information relevance is recognized as a key determinant of information diagnostic for gaining a better understanding of consumers' opinions and their relevance to electronic WOM communications [13]. Online reviews are perceived as relevant when businesses provide information that aligns with consumers' expectations [14].

Previous studies have typically examined factors that contribute to review helpfulness including, text sentiment [15, 16], review depth [17, 18], readability [19-21], novelty [22], credibility [2, 23], specificity [24], reliability [19] and reading enjoyment [25]. However, the research on this topic is quite scattered and inconsistent. Different studies focus on various parts of quality text or use different methods, which makes it hard to get a clear and complete picture. As a result, there is no definitive list of key factors related to information relevance, leading to an incomplete and fragmented understanding of the subject.

Drawing from this observation, this study introduces four key textual features based on information relevance perspective of content novelty, content specificity, content readability and content reliability that potentially contribute to the helpfulness of reviews. These factors represent essential dimensions of information relevance, to enhance quality of online review. Novelty introduces previously unmentioned perspectives, while specificity provides detailed information tailored to the reader's needs. Readability ensures that the information is easily understood, and reliability supports the trustworthiness and accuracy of the content. Therefore, the proposed features can help to better understand how the subjective qualities of reviews impact their helpfulness and influence customer purchasing decisions.

Next, this study adopts a 'threshold' approach to identify helpful reviews by categorizing them based on the number of helpful votes received. The concept of helpful votes threshold provides an innovative way to filter reviews, ensuring that reviews deemed helpful by a larger number of users are given prominence. By setting these thresholds, platforms can prioritize reviews that have resonated with consumers, thereby helping potential buyers make informed decisions more quickly. For instance, reviews exceeding a certain number of helpful votes can be classified as "helpful," allowing the system to highlight feedback that users have found insightful and trustworthy.

Although results did not show a dramatic improvement in classification accuracy, the threshold approach offers practical benefits. It enables a structured and automated system for identifying helpful content, reducing the reliance on manual helpfulness voting and minimizing the risk of helpful reviews being overlooked. The threshold system also aids in understanding how different levels of helpful votes correlate with review helpfulness, providing insights that could guide future improvements in review filtering algorithms. Furthermore, this threshold-based method could serve as a foundation for iterative refinements, where feedback from user interactions helps to adjust the threshold levels dynamically, enhancing the platform's ability to deliver relevant and valuable reviews to consumers.

The rest of this paper is organized as follows. The detail literature reviews of various factors or indicators that contribute to review helpfulness and helpful votes threshold impact on model performance is presented in Section II. Section III introduces methodology that integrates four review text characteristics. Section IV presented results and discussion. Lastly, Section V provides the conclusion of this work.

II. Literature Review

Previous works on review helpfulness have demonstrated association between various review characteristics and review helpfulness and serve as primary source of information over other aspects such as reviewer's identity, metadata and product [26-28]. Research indicates that the content quality, clarity, and emotional tone of the text significantly affect its perceived helpfulness as they enhance the review's credibility and relatability to readers [29]. In addition, text-related features such as length and structure can impact engagement levels, making them more critical than metadata or reviewer-related characteristics, which may not consistently correlate with helpfulness [30, 31]. Therefore, emphasizing text-related characteristics allows for a more direct assessment of review helpfulness.

Previous studies on novelty detection have explored various techniques to identify new and unique information within data. These studies aim to identify new information within a document by employing various approaches tailored to different objectives. Various widely used measurement metrics for novelty detection are utilized across document such as Simple New Word Count, Set Difference, TF-IDF scoring, and Cosine Distance [32, 33]. These methods apply a bag-of-words approach, utilizing word counts within a document. Some novelty measures are derived from probabilistic document models [33-35]. The Simple New Word Count measure, which examines the occurrence of novel words in sentences, has been shown to be as effective as probabilistic document models and other bag-of-words-based methods [33]. Recently, Deep Learning methods such as BERT (Bidirectional Encoder Representations from Transformers) have gained popularity for tasks involving semantic textual similarity [36]. One such method, Sentence-BERT [37], utilizes a pre-trained BERT model to generate context-aware text embeddings, which can be employed to assess the similarity between documents. Sentence-BERT was adapted to calculate the novelty measure in the main analysis, and the analyses were replicated using new word pairs and a revised version of the Simple New Word Count measure [22]. The work also demonstrated review novelty impacts on consumers and businesses.

The discussion on specificity primarily focuses on the sentence level. A common definition of specificity, as used by [38-40], refers to the amount of detail within a sentence. Research on specificity utilizes a broad array of features to indicate sentence specificity. While some studies employ a large collection of features, others may rely on just one. Given the significant variation in feature sets, it is logical to analyze the importance of each feature. Based on current knowledge, the simplest prediction method relies on just one feature which is normalized inverse word frequency (IDF). The sum, average, minimum, and maximum IDF values for all words within a sentence were evaluated, with the maximum IDF value found to be the best indicator of specificity [41]. However, relying on a single feature has its limitations, as the predictor may lose effectiveness in tasks involving multiple topics, due to the significant variation in word distribution across different topics. Speciteller [42] is a popular tool for predicting sentence specificity. It generates 17 features, including sentence characteristics and word representations. Typically, researchers combine Speciteller features with their own custom features to build specificity estimators. For instance, the study by [43] combined Speciteller features with online dialogue features, resulting in a model that outperforms Speciteller in predicting specificity in classroom discussions. In contrast, the work by [24] used a model developed by [38] to assess the specificity of sentences in product reviews. They introduced three new metrics: the percentage of specific sentences in a review, the overall specificity of the review, and the balance between specific and general sentences. This study represents the first attempt to approach the helpfulness prediction problem from sentence specificity perspective.

Review readability, which refers to the ease with which a text can be understood, significantly affects the perceived helpfulness of reviews. Consumers are more likely to find a review helpful if they can easily interpret it. Therefore, higher readability generally facilitates better understanding. Like novelty and specificity, numerous features are used to predict readability. The features commonly employed in readability prediction studies are generally consistent. These features were categorized into semantic and syntactic groups, with an analysis of both the words and sentence structures [44]. Syntactic features include sentence length, average number of characters per word, average number of syllables per word and the percentage of various part-of-speech tags. Semantic features involve the frequency of various 1-, 2-, and 3-word sentences in a review. Many studies have utilized various readability indices, including the Simple Measure of Gobbledygook (SMOG), Automated Readability Index (ARI), Gunning Fog Index (GFI), Flesch–Kincaid Grade Level (FKGL), Coleman–Liau Index (CLI), and Flesch–Kincaid Reading Ease (FKRE) to predict review helpfulness [21, 28, 45] found that a hybrid set of features based on linguistic categories, review metadata, readability, and subjectivity offered the best review predictive performance. Review content features, such as readability, were identified as the most effective predictors of helpfulness [28]. Readability, along with linguistic and psychological features, was utilized to predict the helpfulness of movie reviews [45].

The reliability of online reviews has attracted significant attention in recent years. Various studies have explored the factors that influence the reliability of online reviews, and the methods used to assess and enhance this reliability. The Linguistic Inquiry and Word Count (LIWC) program was used to analyze the proportion of positive and negative words in reviews [46, 47]. It has been discovered that the sentiment of a review, whether positive or negative, along with advice for decision-making and claims of expertise, significantly influences the perceived helpfulness of the review [46]. Additionally, previous studies have highlighted that both sentiment orientation (positive or negative) and the writing style of the review (subjective or objective) are key factors in determining its believability [46, 48]. Consequently, research conducted by [19] leveraged these elements as reliability indicators to assess review helpfulness. The detection of spam reviews is also critical in evaluating online review reliability. Early research focused on identifying spam reviews by detecting copied content [49-51]. Various reliability features, such as Kullback-Leibler divergence, syntactic text characteristics, and review semantics, have been employed to distinguish fake reviews from genuine ones. Additionally, several algorithms have been developed to filter out unreliable reviews [52-53].

The concept of helpful votes threshold in the context of online reviews, is a crucial parameter in binary classification systems. It directly impacts the performance of algorithms that distinguish between helpful and unhelpful reviews by setting a boundary that defines what qualifies as "helpful." An inappropriate threshold can lead to the misclassification of reviews, where helpful reviews are categorized as unhelpful or vice versa, ultimately weakening the model's performance and skewing results.

The choice of classification threshold is essential to improve classification precision. Studies, such as those by Ghose and Ipeirotis [64], have found that a threshold where the ratio of helpful votes to total votes equals 0.6 can significantly enhance classification accuracy for review helpfulness on platforms like Amazon. This threshold value, also adopted by researchers such as Krishnamoorthy [21] and Malik and Hussain [69], minimizes the chances of misclassifying helpful reviews as unhelpful and vice versa, improving the reliability of the helpfulness measure.

III. METHODOLOGY

The section introduces methodology for predicting the helpfulness of online reviews, including the collection of product reviews, review characteristics and the helpfulness of a review as shown in Fig. 1.

A. Data Collection

Many e-commerce platforms provide product or service reviews and relevant data. This study is focused on products available in Amazon.com. The data collected included review rating, review title and text, identification number of a product, user identification number, the time a review is posted, number of helpful votes received by a review and verified purchase of a product. Reviews from the year 2022 until the year 2023 were downloaded for this study [54].

B. Data Preparation

This study utilized a dataset consisting of 9,369 helpful reviews and 9,369 unhelpful reviews sourced from the Beauty, Health, and Personal Care categories on Amazon.com. The dataset is consistent across all helpful votes thresholds employed in this experiment. As many e-commerce platforms provide valuable insights through product and service reviews, this research specifically focuses on products available on Amazon.com.

The collected data included various attributes such as review ratings, review titles, review text, product identification numbers, user identification numbers, timestamps of when the reviews were posted, the number of helpful votes each review received, and whether the purchase was verified. The reviews analyzed were collected from the years 2022 to 2023 [54]. To refine the data before feature extraction, non-English text is filtered out from the dataset. Since text containing URLs and HTML tags might point to a promotional site, or competitors, rows of data with these elements are also removed. In addition, text with emojis and emoticons are also eliminated. Then, the rows of data where the text column is empty or contains fewer than 5 words are excluded, as these offer limited information for potential customers [55]. Besides, data with duplicated text are also omitted. Review duplication can potentially occur in three difference situations [56]:

- Duplicate reviews of the same product with a different user identification number.
- Duplicate reviews from the same user ID but on different products.
- Duplicate reviews from different user IDs on different products.

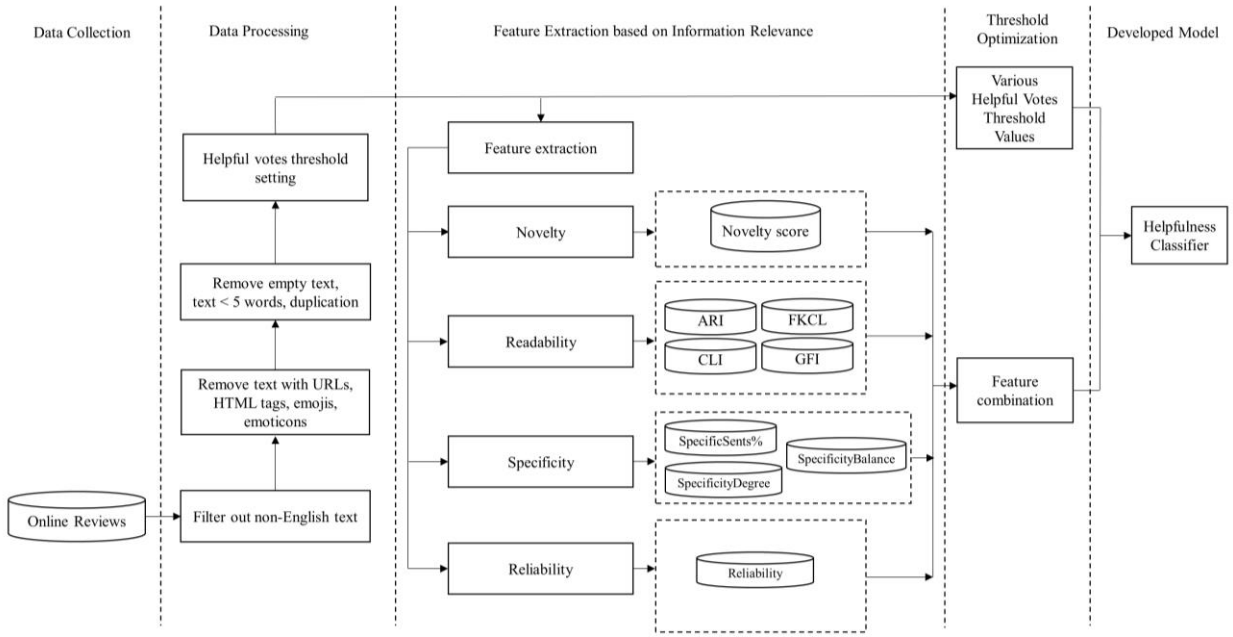

Fig. 1. Research process of this work

In terms of helpful indicator, the study by [57] suggest that highly adopted review helpfulness ratio (number of helpful votes/total number of votes) could lead to highly bias results. Hence, this study utilized the various helpful votes threshold values that is transformed into binary representation.

C. Feature Extraction

1) Content novelty: Some previous studies have described novel information as information that is different from what readers already know or expect [58]. However, most consumers cannot rely solely on their prior knowledge to guide them when making decision in an online environment, especially when purchasing experienced goods [59]. New information and perspective in online reviews may influence consumers to purchase products as it increases consumer's awareness about a product or a service. Consumers highly value knowledge gained from firsthand experience, often spending substantial time and effort searching for and reading reviews to find new information and insights. Therefore, novel information in reviews can be perceived as helpful by consumers.

Empirical evidence indicates that most consumers do not look beyond the first page of search results [60, 61]. Additionally, a consumer survey reveals that most consumers read no more than ten reviews before making a purchase [62]. Hence, content novelty can be defined as the amount of novel information in current review compared with the 10 most recent reviews on a single page. Novelty score for each review is used to determine the amount of novel information in reviews. The method of measuring the amount of novel information in each review is based on method proposed by [22]. First, let r_i represent the focal review and let $C_i = \{r_{i1}, r_{i2}, r_{i3}, \ldots, r_{ij}\}$ represent the comparison set for r_i, where j is the number of reviews in the comparison set (up to 10).

The novelty score for r_i given the comparison set C_i can be expressed as:

$$N(\rho_\iota, X_\iota) = \min_{\rho_{\iota\kappa} \square X_\iota} \left(1 - \chi o \sigma \left(\varepsilon \mu \beta(\rho_\iota), \varepsilon \mu \beta(\rho_{\iota\kappa})\right)\right) \quad (1)$$

Where

- emb(r) represent the embedding of review r using the context-aware representation method.
- $\cos(\text{emb}(r_i), \text{emb}(r_{ik}))$ is the cosine similarity between the embeddings of the focal review r_i and a review r_{ik} in the comparison set C_i.

The final novelty score for the review r_i, considering all possible comparison sets $C_1, C_2, C_3, \ldots, C_n$ is the average of the minimum novelty score across all comparison sets:

$$\Phi \iota \nu \alpha \lambda \ No \varpi \varepsilon \lambda \tau \psi \ \Sigma \chi o \rho \varepsilon \ (\rho_\iota) = \frac{1}{\nu} \sum_{\mu=1}^{\nu} N(\rho_\iota, X_\mu) \quad (2)$$

Where n is the number of different comparison sets considered for r_i and m is the index of the comparison set.

2) Content readability: The readability of a review is another crucial factor that can influence its perceived helpfulness. A review that is highly readable is more likely to be read and voted on by a larger number of users. Readability refers to the ease with which a reader can understand and process a piece of textual information [63]. The readability feature also determines the complexity of any review for the user [45]. Previous research by [64] shows that readability and subjectivity features outperform the lexical features employed by [65]. The work by study [21] shows that the combination of features derived from linguistic categories, readability, review metadata, and subjectivity provide the most accurate predictive performance.

To assess review readability, four grade level readability metrics can be utilized namely (1) Automated Readability Index (ARI), (2) Flesch–Kincaid Grade Level (FKGL), (3) Gunning Fog Index (GFI), and (4) Coleman–Liau Index (CLI) [21].

ARI was selected in this study because it is one of the primary measures used to assess text readability and is less prone to error than other readability measures [66]. The ARI can be calculated using its standard formula:

$$\text{API } \Sigma\chi\text{o}\rho\varepsilon\text{=4.71 } \left(\frac{\nu\upsilon\mu\beta\varepsilon\rho \text{ o}\phi \chi\eta\alpha\rho\alpha\chi\tau\varepsilon\rho\sigma}{\nu\upsilon\mu\beta\varepsilon\rho \text{ o}\phi \omega\text{o}\rho\delta\sigma}\right) +0.5 \left(\frac{\nu\upsilon\mu\beta\varepsilon\rho \text{ o}\phi \omega\text{o}\rho\delta\sigma}{\nu\upsilon\mu\beta\varepsilon\rho \text{ o}\phi \sigma\varepsilon\nu\tau\varepsilon\nu\chi\varepsilon\sigma}\right) - 21.43 \quad (3)$$

By analyzing sentence length, word difficulty, and text cohesion, the FKGL formula measures how challenging readers may find a text. The formula is as follows:

$$\text{K}\Gamma\Lambda\ \Sigma\chi\text{o}\rho\varepsilon = 0.39 \left(\frac{\omega\text{o}\rho\delta\sigma}{\sigma\varepsilon\nu\tau\varepsilon\nu\chi\varepsilon\sigma}\right) - 11.8 \left(\frac{\sigma\psi\lambda\lambda\alpha\beta\lambda\varepsilon\sigma}{\omega\text{o}\rho\delta\sigma}\right) - 15.59 \quad (4)$$

The Gunning Fog Index formula is based on the idea that shorter sentences written in clear, straightforward language receive a better score than longer, more complex sentences. Online reviews that score well on the Gunning Fog Index are likely to be more accessible and comprehensible to a broader audience, enhancing user engagement and improving the overall quality of the review content. The formula is given by

$$\Gamma\Phi\text{I=0.4} \left[\left(\frac{\omega\text{o}\rho\delta\sigma}{\sigma\varepsilon\nu\tau\varepsilon\nu\chi\varepsilon\sigma}\right) + 100 \left(\frac{\chi\text{o}\mu\pi\lambda\varepsilon\xi \ \omega\text{o}\rho\delta\sigma}{\omega\text{o}\rho\delta\sigma}\right)\right] \quad (5)$$

Meanwhile, CLI scores indicate the complexity of a text and are determined using the formula:

$$\text{X}\Lambda\text{I } \Sigma\chi\text{o}\rho\varepsilon\text{=0.0588} \left(\frac{\lambda\varepsilon\tau\tau\varepsilon\rho\sigma}{100 \ \omega\text{o}\rho\delta\sigma}\right) - 0.296 \left(\frac{\sigma\varepsilon\nu\tau\varepsilon\nu\chi\varepsilon\sigma}{100 \ \omega\text{o}\rho\delta\sigma}\right) - 15 \quad (6)$$

3) Content specificity: To assess content specificity in reviews, we adopted a two-step approach that involves calculating a specificity score based on the Normalized Inverse Document Frequency (NIDF) method and then deriving three specific features as outlined in prior research.

Inverse Document Frequency (IDF) is a widely recognized metric that measures the discriminative ability of a term within a document collection. It is defined as the logarithmic ratio of the total number of documents in the collection (n_d) to the number of documents containing the term (known as the term's document frequency, $df(t_i)$), as shown as follow:

$$IDF(t_i) = \log\left(\frac{n_d}{df(t_i)}\right) \quad (7)$$

In this study, we employed the Normalized Inverse Document Frequency (NIDF)The NIDF, defined in equation 8, normalizes with respect to the number of documents not containing the term ($n_d - df(t_i)$) and adds a constant 0.5 to both the numerator and the denominator to moderate extreme values:

$$NIDF(t_i) = \log\left(\frac{n_d - df(t_i) + 0.5}{df(t_i) + 0.5}\right) \quad (8)$$

Commonly used words, such as "the", "and", and "it" are likely to appear in nearly every document and are therefore not particularly discriminative. This lack of discriminative capability is reflected in their low NIDF values. Conversely, terms that occur in only a small number of documents are more useful for distinguishing between documents, resulting in higher NIDF values.

Our assumption is that documents dominated by terms with low NIDF values are less specific than those containing more discriminative terms. Consequently, we define a document specificity score, S_1, as follows:

$$S_1(d) = \frac{1}{l_d} \sum_{t_i \in d} tf(t_i) \cdot \log\left(\frac{n_d - df(t_i) + 0.5}{df(t_i) + 0.5}\right) \quad (9)$$

In this equation, $tf(t_i)$ represents the term frequency of t_i in document d, and l_d denotes the length of document d. The inclusion of l_d in the denominator reduces the impact of varying document lengths on the specificity score.

Following this calculation of the specificity score for each review, we further derived three specific features that have been proposed in previous study [24]. These features aim to assess different aspects of specificity within the context of helpfulness in online reviews.

- SpecificSents% represents the percentage of specific sentences within a review. A sentence is classified as "specific" if its specificity score is 0.5 or higher. This feature is designed to assess whether the number of specific or general sentences in a review affects its perceived helpfulness.
- SpecificityDegree represents the overall specificity of a review. Given the set *P* all sentences in a review and $\sigma(p)$ as the specificity score of a sentence $p \in P$, the specificity degree of the review is defined as:

$$\Sigma\pi\varepsilon\chi\iota\phi\iota\varpi\iota\tau\psi\Delta\varepsilon\gamma\rho\varepsilon\varepsilon = \frac{\sum_{\pi}^{\Pi} \int(\pi)}{|\Pi|} \quad (10)$$

Where |P| is the total number of sentences in the review.

- SpecificityBalance measures the balance between specific and general sentences in a review. Let S be the set of specific sentences (with a specificity degree ≥ 0.5) and G the set of general sentences (with a specificity degree < 0.5) in a review. The specificity balance of a review is calculated as:

$$\Sigma\pi\varepsilon\chi\iota\phi\iota\chi\iota\tau\psi\text{B}\alpha\lambda\alpha\nu\chi\varepsilon = \frac{||\Sigma| - |\Gamma||}{|\Sigma| + |\Gamma|} \quad (11)$$

Where |S| is the number of specific sentences, |G| is the number of general sentences and $||S|-|G||$ represents the absolute difference between these quantities. A value of 0 indicates a perfect balance between specific and general sentences, whereas a value of 1 means the review consists entirely of either specific or general sentences. According to a study by [67], general sentences are vital for high-quality journalism summaries, suggesting that this balance might influence the perception of helpfulness in product reviews as well.

4) Content reliability: Content reliability plays a critical role in determining the trustworthiness of online reviews. For this feature, a binary indicator is used to represent the reliability

of the review, specifically through the presence of a verified purchase. Reviews from verified customers are considered more genuine because they come from individuals who have actually bought and used the product. This authenticity boosts the perceived reliability of the review, as users are more likely to trust feedback from verified buyers, viewing it as a truthful and accurate reflection of the product [68]. By using a binary indicator - where 1 represents a review from a verified purchase and 0 represents a non-verified review—this feature effectively captures the connection between the genuineness of the review and its perceived reliability. Incorporating this binary indicator allows for a more structured evaluation of content reliability, enhancing the accuracy of any model that seeks to assess the trustworthiness and overall value of online reviews.

D. Neural Network Architecture for Helpfulness Prediction

To predict review helpfulness, we developed an Artificial Neural Network (ANN) using a Multi-Layer Perceptron (MLP) classifier. The MLP architecture was carefully tuned to achieve optimal predictive performance by experimenting with the number of hidden neurons and analyzing the network's response to input features.

1) Model architecture: The MLP model receives each review as an input vector of features that capture key aspects related to content quality, readability, novelty, and specificity, which are hypothesized to influence helpfulness. The input layer of the MLP is designed to process these 9 input features (illustrated in Fig. 2), each representing a distinct attribute of the review. This input layer serves as the foundation for the subsequent layers, encoding the feature values as the model begins to learn from the data.

The final MLP model configuration is as follows:

- Input Layer: Accepts a vector of 9 features per review, providing the model with a rich, multi-dimensional view of each review.
- Hidden Layer Size: A single hidden layer with 140 neurons, which was selected as the optimal configuration after experimenting with values of 20, 40, 60, 80, 100, and 140 neurons. This configuration effectively balances complexity and generalization ability.
- Activation Function: The Rectified Linear Unit (ReLU) activation function was used for the hidden layer, providing computational efficiency and mitigating the vanishing gradient problem.
- Solver: The Adam optimizer was employed to train the network. Adam combines the benefits of Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp), ensuring efficient convergence.
- Learning Rate: An initial learning rate of 0.001, which allows the model to learn gradually and converge steadily.
- Epochs: Training was set to a maximum of 2000 iterations, with early stopping to prevent overfitting.
- Random State: A random state of 42 ensures reproducibility across different runs.
- Output Layer: The output layer consists of a single neuron with a sigmoid activation function, which produces a binary output of either 0 or 1 for each review. Here, an output of 1 indicates that the model predicts the review as "helpful" while an output of 0 indicates a prediction of "not helpful".

The optimal configuration of 140 hidden neurons, based on its predictive accuracy, provided the best balance between model complexity and performance. By systematically testing different neuron counts, we identified this structure as the most suitable for the task of review helpfulness prediction.

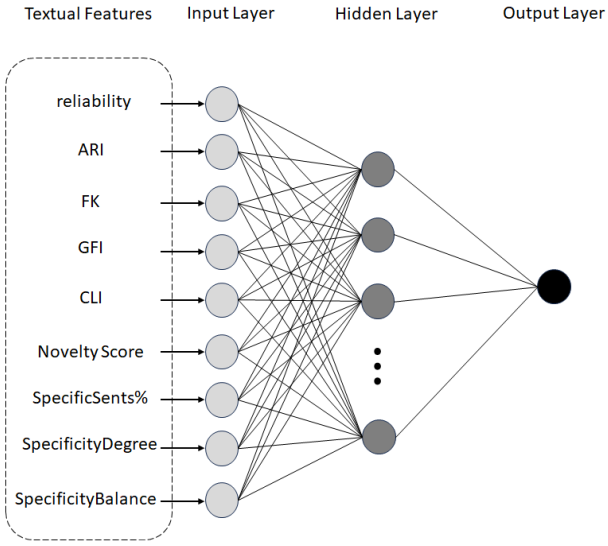

Fig. 2. Layout of MLP model.

2) Performance metrics: To evaluate the effectiveness of our model in classifying online reviews as helpful or unhelpful, we utilized several performance metrics: accuracy, precision, recall, and F1 score. Each of these metrics provides insight into different aspects of model performance, particularly in the context of user-generated content where the classification of reviews can significantly impact consumer decision-making.

Accuracy measures the overall correctness of the model's predictions. It is defined as the ratio of the number of correct predictions to the total number of predictions made. In the context of online reviews, it can be expressed mathematically as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Where

- TP is True Positives (number of helpful reviews correctly classified as helpful).
- TN is True Negatives (number of unhelpful reviews correctly classified as unhelpful).
- FP is False Positives (number of unhelpful reviews incorrectly classified as helpful).

- FN is False Negatives (number of helpful reviews incorrectly classified as unhelpful).

Precision quantifies the accuracy of the positive predictions made by the model, focusing specifically on how many of the predicted helpful reviews are actually helpful. This metric is particularly important in online reviews, where consumers benefit from identifying truly helpful feedback. Precision is calculated as follows:

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

Recall, also known as sensitivity, measures the model's ability to identify all relevant instances of helpful reviews. It assesses how many of the actual helpful reviews were correctly identified by the model. Recall is computed using the formula:

$$Precision = \frac{TP}{TP+FN} \quad (14)$$

The F1 score provides a balance between precision and recall, offering a single metric that captures both aspects of the model's performance. It is especially useful in scenarios where there is an uneven class distribution, such as when the number of helpful reviews significantly differs from unhelpful reviews. The F1 score is calculated as:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

In the context of online reviews, these metrics allow us to assess how well our model performs in distinguishing between helpful and unhelpful reviews. A high accuracy indicates that the model correctly identifies most reviews, while high precision ensures that consumers can trust the reviews labeled as helpful. Additionally, high recall signifies that the model successfully captures most helpful reviews, which is crucial for users seeking reliable information. The F1 score serves as a comprehensive measure, ensuring a balanced consideration of both precision and recall, which is vital in enhancing the overall consumer experience in online platforms.

E. Optimization of Helpful Votes Threshold Values

The use of Amazon.com's publicly available dataset, which contains only helpful votes, provides a reliable measure of review quality based on user interaction. Given the time constraints of this study, this dataset allowed for an efficient and effective investigation into the impact of helpfulness thresholds. Future research could expand this investigation by incorporating datasets that include total votes for comparison.

In this study, we examine various thresholds for helpful votes, ranging from more than 2 to more than 12, using intervals of 2 votes. Reviews with helpful votes exceeding the specified threshold are classified as helpful, while those with votes below the threshold are deemed unhelpful. Table I summarizes these thresholds.

The experiment stops at the helpful votes threshold of more than 14 because the data available for that threshold is not consistent with the data used for the previous thresholds (more than 2, 4, 6, 8, 10, and 12). For accurate model performance comparisons, it is crucial that the number of data points remains the same across all thresholds. This consistency ensures that any observed differences in model performance can be attributed to the thresholds themselves rather than variations in data quantity.

TABLE I. HELPFUL AND UNHELPFUL REVIEWS BY HELPFUL VOTES THRESHOLD

Helpful Votes Threshold	Helpful Reviews	Unhelpful Reviews
More than 2	> 2	≤ 2
More than 4	> 4	≤ 4
More than 6	> 6	≤ 6
More than 8	> 8	≤ 8
More than 10	> 10	≤ 10
More than 12	> 12	≤ 12

IV. RESULTS AND DISCUSSION

Fig. 2 illustrates the relationship between test accuracy and the number of hidden neurons in a neural network across different helpful vote thresholds. The x-axis represents the number of hidden neurons, ranging from 20 to 140, while the y-axis displays test accuracy as a percentage. Each line corresponds to a different threshold for the number of helpful votes, ranging from "> 2" to "> 12," offering insights into how the model performs with varying thresholds.

A general trend shows that test accuracy improves slightly with an increasing number of hidden neurons, though the gains are more pronounced for certain thresholds. The performance tends to stabilize around 100–140 hidden neurons for most thresholds, but the overall accuracy is highly dependent on the threshold used.

The helpful vote thresholds have a significant impact on performance. Lower thresholds, particularly "> 2" and "> 4", exhibit the lowest performance, with test accuracies ranging from 66% to 68%. This indicates that when the model includes reviews with very few helpful votes, it struggles to make accurate predictions. In contrast, moderate thresholds such as "> 6" and "> 8" show improved accuracies, ranging between 70% and 72%, but still do not reach the highest performance levels.

The highest test accuracies, around 73–74%, are achieved with higher thresholds such as "> 10" and "> 12". These thresholds indicate that when the model focuses on reviews with a greater number of helpful votes, it is able to generalize more effectively and perform better. Notably, the highest overall accuracy, approximately 74%, occurs at a threshold of "> 12" with 80 hidden neurons. This suggests that reviews with many helpful votes provide the model with more reliable data for classification, possibly due to clearer distinctions between helpful and unhelpful reviews in these subsets.

In summary, higher thresholds for helpful votes (such as "> 10" and "> 12") combined with around 80–100 hidden neurons offer the best classification performance, while lower thresholds lead to poorer model accuracy. The moderate thresholds (Fig. 3) provide a middle ground, but ultimately, the model benefits most from training on reviews with a larger number of helpful votes, which may contain clearer patterns for the network to learn from.

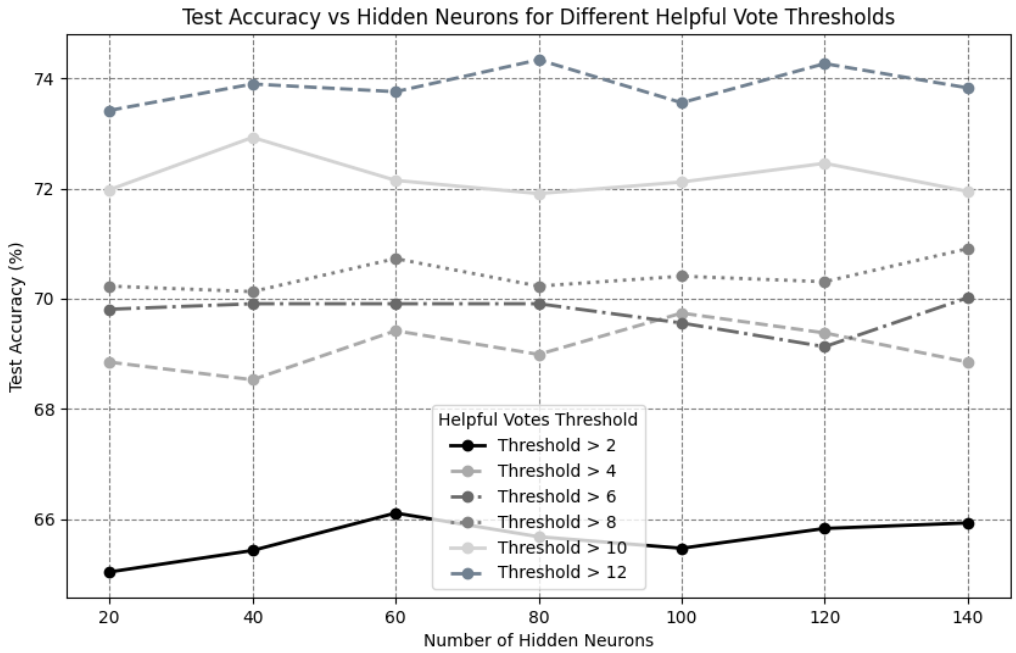

Fig. 3. Results of test accuracy with various helpful votes threshold and number of hidden neurons.

Table II presents model performance across various helpful vote thresholds and hidden neuron configurations, with metrics such as accuracy, precision, recall, and F1 score. These metrics provide a comprehensive evaluation of how well the model classifies reviews as helpful or unhelpful.

TABLE II. THE TEST ACCURACY, PRECISION, RECALL, AND FI SCORE FOR VARIOUS HELPFUL VOTES THRESHOLD AND NUMBER OF HIDDEN NEURONS

Helpful Votes Threshold	Hidden Neuron	Accuracy (%)	Precision	Recall	F1 Score
> 2	60	66.11	0.65	0.70	0.67
> 4	60	69.42	0.68	0.72	0.70
> 6	140	70.02	0.69	0.72	0.71
> 8	140	70.91	0.72	0.69	0.70
> 10	40	72.93	0.72	0.74	0.73
> 12	80	74.34	0.75	0.73	0.74

The model's accuracy improves steadily as the threshold for helpful votes increases. For the lowest thresholds, "> 2" and "> 4", the model achieves accuracy scores of 66.11% and 69.42%, respectively. As the threshold increases to "> 6" and "> 8", accuracy rises to 70.02% and 70.91%. The highest accuracy, 74.34%, is achieved with the "> 12" threshold and 80 hidden neurons. This suggests that reviews with a higher number of helpful votes are easier for the model to classify, possibly due to clearer patterns of helpfulness in more highly voted reviews.

Precision generally improves as the threshold increases, indicating the model's growing ability to correctly identify helpful reviews as the threshold rises. For the "> 2" and "> 4" thresholds, precision starts at 0.65 and 0.68, respectively. It peaks at 0.75 for the "> 12" threshold, demonstrating that the model is better at reducing false positives (i.e., labeling unhelpful reviews as helpful) when working with reviews that have garnered more helpful votes.

The recall metric, which measures the model's ability to correctly identify all helpful reviews, is relatively stable across different thresholds, ranging from 0.69 to 0.74. Interestingly, the highest recall value of 0.74 is achieved at the "> 10" threshold, slightly higher than the recall at the "> 12" threshold (0.73). This suggests that the model is not significantly more likely to miss helpful reviews as the threshold increases.

The F1 score, which balances precision and recall, follows a similar trend to accuracy. It starts at 0.67 for the "> 2" threshold and gradually increases to 0.74 for the "> 12" threshold. This indicates that as the threshold rises, the model becomes better at balancing false positives and false negatives. The highest F1 score (0.74) at the "> 12" threshold demonstrates the model's overall strongest performance with this higher threshold.

The model's performance improves consistently as the threshold for helpful votes increases. Higher thresholds, such as "> 10" and "> 12," yield the best results in terms of accuracy, precision, and F1 score, showing that reviews with more helpful votes provide better training data for classifying helpfulness. In particular, the threshold "> 12" paired with 80 hidden neurons achieves the highest overall accuracy (74.34%) and F1 score (0.74), making it the optimal configuration in this context. However, recall remains relatively stable across thresholds, indicating that the model consistently identifies a high proportion of helpful reviews regardless of the threshold. This analysis suggests that setting a higher helpful votes threshold allows the model to focus on more reliable data, leading to better classification performance. Overall, the findings indicate a robust performance in the model's classification capabilities, highlighting its potential for assisting users in identifying helpful content amidst a large volume of reviews.

The model's performance may be impacted by the presence of both search and experience products within the beauty, health, and personal care category. Search and experience products have distinct characteristics that influence how consumers evaluate reviews. For search products, features like skin type compatibility, ingredients, and fragrance are objective and easy to compare [70], [71], [18], leading consumers to generally agree on these qualities. Consequently, reviews for search products often need a higher number of helpful votes to meet consumer expectations for helpfulness, as they offer limited new insights [72].

On the other hand, experience products involve more subjective aspects, with consumers forming varied opinions based on personal experience [71]. Reviews for these products tend to provide unique, valuable perspectives [73] that consumers find helpful even if they receive fewer helpful votes. This difference means that a dataset containing both types of products may create diversity in review patterns that affect the model's ability to generalize and accurately predict helpfulness. As a result, the model might struggle to perform optimally compared to models trained exclusively on either search or experience product data.

V. Conclusion

This study presents a research process aimed at enhancing the classification of helpful reviews on online platforms, ultimately improving the experience for consumers navigating through vast amounts of user-generated content. The findings demonstrate that the performance of our model significantly improves as the threshold for helpful votes increases, particularly when combined with an optimal number of hidden neurons. The highest test accuracy and F1 score were achieved with thresholds of "> 10" and "> 12," suggesting that reviews receiving greater helpful votes provide clearer patterns for the model to learn from.

While the model achieved balanced F1 scores and demonstrated robustness in distinguishing between helpful and unhelpful reviews, certain limitations emerged, particularly in handling both search and experience products within the beauty, health, and personal care domains. The varied characteristics between these product types - where search products offer more objective qualities, and experience products depend heavily on subjective assessments - introduced challenges in model generalization. This differentiation emphasizes the need for adaptive models that can account for the unique features of each product type, potentially by utilizing additional domain-specific content indicators.

Overall, the findings underscore the value of content-based features in predicting review helpfulness and provide a foundation for future advancements in online review classification systems. By identifying and prioritizing content that is likely to aid consumer decision-making, e-commerce platforms can enhance the user experience, foster trust, and promote customer engagement. This work contributes to the ongoing effort to make consumer review systems more effective and efficient in delivering relevant, trustworthy information to users. Furthermore, this research flow can be integrated into existing e-commerce platforms by enhancing the review sorting mechanisms to prioritize helpful reviews based on consumer preferences. For example, platforms could implement an algorithm that displays reviews with high helpfulness scores at the top of product pages, facilitating quicker decision-making for consumers. Additionally, the framework could offer visual indicators for reviews deemed most helpful, improving user engagement with content that meets their specific needs. By integrating our framework, e-commerce platforms could not only improve the quality of information presented to consumers but also foster greater trust and satisfaction, ultimately leading to enhanced purchasing decisions.

While this study has demonstrated significant advancements in predicting the helpfulness of online reviews using information relevance theory and Artificial Neural Networks (ANN), certain limitations must be acknowledged. Firstly, the dataset utilized focuses exclusively on reviews from a single platform (Amazon) within specific product categories, which may limit the generalizability of the findings to other e-commerce platforms or product types. Additionally, the binary classification approach based on helpful votes thresholds may not fully capture the nuanced perceptions of review helpfulness among diverse consumer groups. Another limitation is the potential bias introduced by the predominance of English-language reviews, which excludes multilingual perspectives and insights. Lastly, the variability in consumer behavior across search and experience products introduces challenges in model generalization, underscoring the need for adaptive or hybrid approaches that consider product-specific factors for improved predictive accuracy. Future research should address these limitations to enhance the robustness and applicability of the proposed model.

VI. Future Work

Future research could explore several directions. First, it would be valuable to explore whether the conclusions drawn from this study apply to data from other online marketplaces beyond Amazon.com, which could broaden the relevance of our findings. Second, expanding the dataset reviews spanning a broader range of years to assess the framework's robustness over time and capture evolving trends in review helpfulness. Third, integrating other advanced natural language processing techniques, such as sentiment analysis and topic modeling, could enhance the model's ability to capture intricate aspects of helpfulness that vary across review types. Additionally, employing adaptive thresholding, which dynamically adjusts based on product type or user interaction data, might yield a more tailored approach to identifying helpful reviews. Finally, investigating the potential of hybrid models that combine content features with reviewer and product metadata could offer a more comprehensive framework for evaluating review helpfulness across various platforms and consumer needs.

References

[1] N. Sahoo, C. Dellarocas, and S. Srinivasan, "The impact of online product reviews on product returns", Inf. Sys. Res., vol. 29, no. 3, pp. 723–738, 2018.

[2] K. Pooja, and P. Upadhyaya, "What makes an online review credible? A systematic review of the literature and future research directions," Manag. Rev. Q., vol. 74, pp. 627–659, 2024.

[3] Y. Wu, L. Chen, E. W. T. Ngai, and P. Wu, "Stimulating positive reviews by combining financial and compassionate incentives," Internet Res., https://doi.org/10.1108/INTR-01-2023-0062, 2024.

[4] J. Li, X. Xu, and E. W. T. Ngai, "How review content, sentiment and helpfulness votes jointly affect trust of reviews and attitude", Internet Res., https://doi.org/10.1108/INTR-01-2023-0025, 2024.

[5] J. Luo, Y. Zhang, Y. Guo, and J. Zhang, "A novel method based on knowledge adoption model and non-kernel SVM for predicting the helpfulness of online reviews," J. Oper. Res. Soc., vol. 75, no. 6, pp. 1205-1222, 2024.

[6] X. Liu, G. A. Wang, W. Fan, and Z. Zhang, "Finding useful solutions in online knowledge communities: A theory-driven design and multilevel analysis," Inf. Syst. Res., vol. 31, no. 3, 2020.

[7] A. H. Huang, K. Chen, D. C. Yen, and T. P. Tran, "A study of factors that contribute to online review helpfulness," Comput. Human Behav., vol. 48, pp. 17–27, 2015.

[8] C. Cheung, M. Lee, and N. Rabjohn, "The impact of electronic word-of-mouth – The adoption of online opinions in online customer communities," Internet Res., vol. 18, no. 3, pp. 229–247, 2008.

[9] S. W. Sussman, and W. S. Siegal, "Informational influence in organizations: An integrated approach to knowledge adoption," Inf. Syst. Res., vol. 14, no. 1, pp. 47-65, 2003.

[10] S. Watts, G. Shankaranarayanan, and A. Even, "Data quality assessment in context: a cognitive perspective," Decis. Support Syst., vol. 48, no. 1, pp. 202–211, 2009.

[11] Y. C. Chen, R. A. Shang, and M. J. Li, "The effects of perceived relevance of travel blogs' content on the behavioral intention to visit a tourist destination," Comput. Hum. Behav., vol. 30, pp. 787–799, 2014.

[12] A. L. Jepsen, "Factors affecting consumer use of the Internet for information search," J. Interact. Mark., vol. 21, no. 3, pp. 21–34, 2007.

[13] R. Filieri, "What makes online reviews helpful? A diagnosticity-adoption Framework to explain informational and normative influences in e-WOM," Journal of Business Research, 68(6), 1261–1270, 2015.

[14] R. Filieri, and F. McLeay, "E-WOM and accommodation: an analysis of the factors that influence travelers' adoption of information from online reviews," J. Travel Res., vol. 53, no. 1, pp. 44-57, 2014.

[15] J. L. Nicolau, Z. Xiang, and D. Wang, "Daily online review sentiment and hotel performance," Int. J. Contemp. Hosp. Manag., vol. 36, no. 3, pp. 790 – 811, 2024.

[16] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews", Comput. Sci. Rev., vol. 41, pp. 1- 17, 2021.

[17] Y. -H. Cheng, and H. -Y. Ho, "Social influence's impact on reader perceptions of online reviews," J. Bus. Res., vol. 68, no. 4, pp. 883-887, 2015.

[18] S. M. Mudambi, and D. Schuff, "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," MIS Quarterly, vol. 34, no. 1, pp. 185-200, 2010.

[19] Y. Meng. N. Yang, Z. Qian, and G. Zhang, "What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values," J. Theor. Appl. Electron. Commer. Res., vol. 16, no. 3, pp. 466-490, 2020.

[20] N. S. B. C. Lah, A. R. B. C. Hussin, and H. M. Dahlan, "A concept-level approach in analyzing review readership for E-Commerce persuasive recommendation," *Proceedings of the International Conference on Research and Innovation in Information Systems*, Langkawi, Malaysia, pp. 1-5, 2017.

[21] S. Krishnamoorthy, "Linguistic features for review helpfulness prediction," Expert Syst. Appl., vol. 42, pp. 3751-3759, 2015.

[22] D. Y. Ozdemir, "Essays on Novelty in Online Reviews," [Doctoral Thesis, The University of Texas at Dallas], 2023.

[23] A. G. Mumuni, K. O'Reilly, A. MacMillan, S. Cowley, and B. Kelley, "Online Product Review Impact: The Relative Effects of Review Credibility and Review Relevance," J. Internet Commer., vol. 19, no. 2, 2020.

[24] B. Lima, and T. Nogueira, "Novel features based on sentence specificity for helpfulness prediction of online reviews," 8th Brazilian Conference on Intelligent Systems (BRACIS), pp. 84–89, 2019.

[25] Z. Liu, and S. Park, "What makes a useful online review? Implication for travel product websites," Tour. Manag., vol. 47, pp. 140-151, 2015.

[26] S. J. S. Quaderi, and K. D. Varathan, "Identification of significant features and machine learning technique in predicting helpful reviews," PeerJ Comput. Sci., vol. 10, e1745, 2024.

[27] J. Kong and C. Luo, "Do cultural orientations moderate the effect of online review features on review helpfulness? A case study of online movie reviews," J. Retail. Consum. Serv., vol. 73, 2023.

[28] M. S. I. Malik, "Predicting users' review helpfulness: the role of significant review and reviewer characteristics," Soft Comput., vol 24, no. 18, pp. 13913-13928, 2020.

[29] M. Akgül, and A. R. Montazemi, "Online Review Helpfulness: A Literature Review," In M. Khosrow-Pour, D.B.A. (Ed.), Encyclopedia of Information Science and Technology, Sixth Edition, Advance online publication, 2025. https://doi.org/10.4018/978-1-6684-7366-5.ch055

[30] B. Ganguly, P. Sengupta and B. Biswas, "What are the significant determinants of helpfulness of online review? An exploration across product-types," J. Retail. Consum. Serv., vol. 78, 103748, 2024.

[31] X. Li, Q. Li, and J. Kim, "A Review Helpfulness Modeling Mechanism for Online E-commerce: Multi-Channel CNN End-to-End Approach," Appl. Artif. Intell., vol. 37, no. 1, 2023.

[32] T. Ghosal, T. Saikh, T. Biswas, A. Ekbal, and P. Bhattacharyya, "Novelty Detection: A Perspective from Natural Language Processing," Comput. Linguist., vol.48, no. 1, pp. 77–117, 2022.

[33] J. Allan, C. Wade and A. Bolivar, "Retrieval and novelty detection at the sentence level," In Proc. 26th Annual Internat. ACM SIGIR Conf. Res. and Development Inform. Retreival, pp 314–321, 2003.

[34] Zhang, J., Z. Ghahramani, and Y. Yang (2004). A probabilistic model for online document clustering with application to novelty detection. Adv. Neural Inform. Processing Systems 17.

[35] Zhang, Y., J. Callan, and T. Minka (2002). Novelty and redundancy detection in adaptive filtering. In Proc. 25th Annual Internat. ACMSIGIR Conf. Res. and Development Inform. Retrieval, pp. 81–88.

[36] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference.

[37] Reimers, N. and I. Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[38] W-. J. Ko, G. Durrett, and J. J. Li, "Domain agnostic real-valued specificity prediction," in 33rd AAAI Conf. Artif. Intell., 2019.

[39] A. Louis, and A. Nenkova, "Automatic identification of general and specific sentences by leveraging discourse annotations," In Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 605–613, 2011.

[40] A. Louis, and A. Nenkova, "A corpus of general and specific sentences from news," In Proceedings of the Eighth International Conference on Language Resources and Evaluation, pp. 1818–1821, 2012.

[41] R. Zhang, J. Guo, Y. Fan, Y. Lan, J. Xu, and X. Cheng "Learning to Control the Specificity in Neural Response Generation," In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, pp. 1108–1117, 2018.

[42] J. J. Li, and A. Nenkova, "Fast and accurate prediction of sentence specificity," In Proc of 29th AAAI Conf. Artif. Intell., pp 2281–2287, 2015.

[43] L. Lugini, and D. Litman, "Predicting Specificity in Classroom Discussion," In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 52–61, 2017.

[44] X. Liu, W. B. Croft, P. Oh, and D. Hart, "Automatic recognition of reading levels from user queries," In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.548-549, 2004.

[45] M. S. Khan, A. Rizwan, M. S. Faisal, T. Ahmad, M. S. Khan, and G. Atteia, "Identification of Review Helpfulness Using Novel Textual and Language-Context Features," Mathematics, vol. 10, 3260, 2022.

[46] Chua, A.Y.; Banerjee, S. Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. Comput. Hum. Behav. 2016, 54, 547–554.

[47] J. W. Pennebaker, R. J. Booth, and M. E. Francis, “Linguistic inquiry and word count: LIWC”, 2007. www.liwc.net

[48] Rausser, G.C.; Simon, L.; Zhao, J. Rational exaggeration and counter-exaggeration in information aggregation games. Econ. Theory 2015, 59, 109–146.

[49] C. L. Lai, K. Q. Xu, R. Y. K. Lau, Y. Li and D. Song, "High-Order Concept Associations Mining and Inferential Language Modeling for Online Review Spam Detection," 2010 IEEE International Conference on Data Mining Workshops, Sydney, NSW, Australia, pp. 1120-1127, 2010a.

[50] C. L. Lai, K. Q. Xu, R. Y. K. Lau, Y. Li and L. Jing, "Toward a Language Modeling Approach for Consumer Review Spam Detection," 2010 IEEE 7th International Conference on E-Business Engineering, Shanghai, China, pp. 1-8, 2010b.

[51] R. Y. K. Lau, S. Y. Liao, R. C. -W. Kwok, K. Xu, Y Xia, and Y. Li, “Text mining and probabilistic language modeling for online review spam detection,” ACM Trans. Manage. Inf. Syst., vol. 2, no. 4, 2012.

[52] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “What Yelp Fake Review Filter Might Be Doing?,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 7, no. 1, pp. 409-418, 2013.

[53] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, ‘‘Spotting fake reviews via collective positive-unlabeled learning,’’ in Proc. IEEE Int. Conf. Data Mining, pp. 899–904, 2014.

[54] McAuley Lab, Amazon Reviews’23 [Data set], 2024. https://amazon-reviews-2023.github.io/

[55] S. Xiao, G. Chen, C. Zhang, and X. Li, “Complementary or Substitutive? A Novel Deep Learning Method to Leverage Text-image Interactions for Multimodal Review Helpfulness Prediction,” Expert Syst. Appl., vol. 208, 118138, 2022.

[56] N. Jindal, and B. Liu, “Analyzing and Detecting Review Spam”, in Proc IEEE Int. Conf. Data Min., pp. 547-552, 2007.

[57] X. Guo, G. Chen, C. Wang, Q. Wei, and Z. Zhang, “Calibration of Voting-Based Helpfulness Measurement for Online Reviews: An Iterative Bayesian Probability Approach,” INFORMS J. Comput., vol. 33, no. 1, pp. 246-261, 2021.

[58] M. T. van Kesteren, D. J. Ruiter, G. Fernández, and R. N. Henson, “How schema and novelty augment memory formation,” Trends Neurosci., vol. 35, no. 4, pp. 211-219, 2012.

[59] W. Zhu, J. Mou, and M. Benyoucef, ‘Exploring purchase intention in cross-border E-commerce: A three stage model,” J. Retail. Consum. Serv., vol. 51, pp. 320–330, 2019.

[60] D. R. Fesenmaier, Z. Xiang, B. Pan, and R. Law, “A framework of search engine use for travel planning”, J. Travel Res., vol. 50, no. 6, pp. 587-601, 2011.

[61] B. J. Jansen, and A. Spink, “How are we searching the world wide web? A comparison of nine search engine transaction logs,” Inf. Process. Manag., vol. 42, no. 1, pp. 248-263, 2006.

[62] S. Paget, “Local Consumer Review Survey 2024: Trends, Behaviors, and Platforms Explored,” Accessed 9 August 2024, https://www.brightlocal.com/research/local-consumer-review-survey/.

[63] I. Raoofpanah, C. Zamudio, and C. Groening, “Review reader segmentation based on the heterogeneous impacts of review and reviewer attributes on review helpfulness: a study involving ZIP code data,” J. Retailing Consum. Serv., vol. 72, 103300, 2023.

[64] A. Ghose, and P. Ipeirotis, “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics,” IEEE Trans. Knowl. Data Eng., vol. 23, pp. 1498–1512, 2011.

[65] Z. Zhang, and B. Varadarajan, “Utility scoring of product reviews,” In Proc of the 15th ACM Int. Conf. Inf. Knowl. Manag., pp. 51–57, 2006.

[66] N. Hu, I. Bose, Y. Gao, and L. Liu, "Manipulation in digital word-of-mouth: A reality check for book reviews", Decis. Support Syst., vol. 50, no. 3, pp. 627-635, 2011.

[67] Annie Louis and Ani Nenkova. Text Specificity and Impact on Quality of News Summaries. In Proceedings of the Workshop on Monolingual Text-To-Text Generation, pages 34–42, 2011.

[68] M. J. Kim, K. K.-C., Park, M. Mariani, and S. F. Wamba, “Investigating reviewers' intentions to post fake vs. authentic reviews based on behavioral linguistic features,” Technol. Forecast. Soc. Change, vol. 198, pp. 122971, 2024.

[69] M.S.I. Malik, and A. Hussain, “An analysis of review content and reviewer variables that contribute to review helpfulness,” Inf. Process. Manage., vol. 54, no. 1, pp. 88–104, 2018.

[70] M. J. Kim, K. K.-C., Park, M. Mariani, and S. F. Wamba, “Investigating reviewers' intentions to post fake vs. authentic reviews based on behavioral linguistic features,” *Technol. Forecast. Soc. Change*, vol. 198, pp. 122971, 2024.

[71] L. Huang, C.-H. Tan, W. Ke, and K.-K. Wei, “Comprehension and assessment of product reviews: a review-product congruity proposition”, *J. Manag. Inf. Syst.*, vol. 30, no. 3, pp. 311–343, 2013.

[72] S. Xiao, G. Chen, C. Zhang, and X. Li, “Complementary or Substitutive? A Novel Deep Learning Method to Leverage Text-image Interactions for Multimodal Review Helpfulness Prediction,” *Expert Syst. Appl.*, vol. 208, 118138, 2022.

[73] Y. Ma, Z. Xiang, Q. Du, and W. Fan, “Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning,” *Int. J. Hosp. Manag.*, vol. 71, pp. 120-131, 2018.