# End-to-End Assessment of Product Review Helpfulness Using Subjective and Objective Information

Yuta Nakajima[1,*], Michal Ptaszynski[1,*] and Fumito Masui[1]

[1]Kitami Institute of Technology, 165 koencho, kitami, Hokkaido, Japan 090-8507

**Abstract**

With the spread of the internet, the volume of reviews on services, shopping, and word-of-mouth websites has grown annually, becoming a significant source of information for decision-making. However, not all reviews are equally useful to potential buyers. A certain number of reviews are low-value or spam. Therefore, presenting only highly useful reviews to users would support more effective and efficient decision-making. In this study, we address this issue by analyzing the features of useful reviews and proposing a method to support user decision-making. First, after surveying existing research on review helpfulness classification, we propose a scoring method that focuses on the amount of information in a document (text volume) as a key feature of useful reviews. We also examined the concept of helpfulness of reviews in detail, and created a dataset containing reviews with annotations of subjectively perceived helpfulness, and objective features related to it. In a binary classification experiment using Transformers-based models to categorize review helpfulness, we obtained strong results, with an F1 score exceeding 80%. Furthermore, to show users which parts of a review are useful, we built a multi-label classification model to automatically extract the features of helpfulness. This model demonstrated its ability to effectively capture the characteristics of useful reviews, achieving an F1 score of over 80% for four of the core helpfulness-related features defined in this research.

**Keywords**

Online Product Reviews, Review helpfulness Prediction, Multi-label Classification, Transformers, Lexical Density

## 1. Introduction

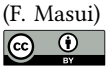
In recent years, with the spread of the internet, user reviews are posted across all types of media, such as online shopping and accommodation booking sites, broadly influencing the sharing of word-of-mouth information and user decision-making. However, users must browse through a massive number of reviews for products they are considering, which requires a great deal of their time and effort. When the number of reviews is too large to read, users often proceed with their purchase consideration without reading most of the reviews. However, unread reviews may contain valuable information, while those read first frequently contain redundant or low-quality information that is not helpful for decision-making. Moreover, there is an increasing number of spam reviews (reviews unrelated to the product, reviews that serve as advertisements for other products, etc.) and fake reviews (reviews written by individuals who have not purchased or used the product, either of their own volition or by request, to excessively praise or criticize the product). Therefore, it is important for users considering a product purchase to read only high-quality, useful reviews, creating a need for the development of methods to determine review helpfulness.

In this study, we investigate the characteristics of highly useful reviews and establish a definition to determine not only whether a review is useful, but also which aspects of it are considered useful. Furthermore, similar to previous research, we aim to automatically select highly useful reviews from a large number, thereby providing users with a metric for decision-making during purchases and supporting them without requiring them to read all the reviews.

The remainder of this paper is structured as follows. We begin by reviewing related work in Section 2. Section 3 details our complete methodology, including our framework for defining helpfulness, the novel 'lds' score for data sampling, and the annotation process used to create our dataset. In Section 4,

*LaCATODA 2026: The 10th Linguistic and Cognitive Approaches to Dialog Agents Workshop at the 40th AAAI Conference*

m3245300181@std.kitami-it.ac.jp (Y. Nakajima); michal@mail.kitami-it.ac.jp (M. Ptaszynski); f-masui@mail.kitami-it.ac.jp (F. Masui)



we present our main experimental results, covering both the binary helpfulness classification and our primary multi-label feature classification task. Following this, we discuss the broader implications of our findings in Section 5. Finally, we summarize our contributions and outline directions for future research in Section 6.

## 2. Related Work

Research on review helpfulness prediction has a long history, beginning with models that used engineered features from review content and metadata, to a more recent studies which shifted towards deep learning models that learn features directly from text.

Kim et al. [1] pioneered this area by using a Support Vector Regression (SVR) model with features like reviewer history, review length, and unigrams. Our work differs by employing a multi-label classification framework for explainability rather than regression, and by using end-to-end Transformer models to learn semantic representations, which reduces the need for manual feature engineering.

Mudambi and Schuff [2] confirmed that review length together with explicitly expressed rating are strong predictors of helpfulness, particularly when considering the product type (search vs. experience). Building on their findings, our research focuses more on the semantic content of the review. We also introduce the ‘lds‘ score, a metric that considers lexical diversity while penalizing short texts, as a more sophisticated measure of informativeness than length alone.

Zhang and Tran [3] proposed an efficient linear model using review length to address the "cold-start" problem for new reviews. We advance this by using non-linear Transformer architectures that better interpret complex semantics. Furthermore, our multi-label framework provides an explainable output identifying *why* a review is useful, being an improvement over a single, uninterpretable ranking score.

Pan and Zhang [4] demonstrated the importance of reviewer metadata, such as post history, in predicting helpfulness. In contrast, our work deliberately focuses only on the review’s content. This design choice makes our model more universally applicable, especially on platforms where reviewer metadata is unavailable, and relies on deep learning to extract all necessary signals from the text itself.

Sasaki et al. [5] established eight criteria for helpfulness in Japanese reviews and applied an SVM to morphological features. Our work adopts their definitional approach but validates it on a larger, crowd-sourced dataset and uses more advanced deep learning models for classification.

The shift to deep learning also in this area of research is represented by Qu et al. [6], who used a Convolutional Neural Network (CNN) to learn features from word embeddings. Our study represents the next methodological step by employing Transformer architectures. The self-attention mechanism in Transformers is better suited to modeling the long-range contextual dependencies within a review compared to the local feature detection of CNNs.

Zhang and Lin [7] addressed multilingual helpfulness prediction for English and German reviews using a combination of language-dependent and independent features. While their work focused on multilingual breadth, our research on the other hand focuses on monolingual depth. We propose a detailed set of six helpfulness criteria for Japanese and an explainable multi-label detection model.

Sun et al. [8] studied the concept of "informativeness," highlighting its complexity and dependence on product type. Our work improves on their approach by proposing the ‘lds‘ score as a concrete metric for informativeness and uses it as a strategic tool to sample high-quality data for further annotation.

Saumya et al. [9] also applied a CNN to predict a continuous helpfulness score. Our work advances this by using a more powerful Transformer architecture and, more importantly, by formulating the problem as an explainable multi-label classification task rather than a regression task that produces a single, unexplainable score.

Malik [10] found that review-specific content features were the strongest predictors of helpfulness, more so than reviewer or product-type features. Our work validates this finding by focusing exclusively on content. We demonstrate that state-of-the-art language models can extract a rich set of predictive signals from the review text alone, making our system independent of external metadata.

Soda et al. [11] aimed to show users not just whether a customer review is useful or not, but in what

way. They proposed seven perspectives for evaluating Japanese reviews' helpfulness and automated three of them. Our multi-label classification approach shares this goal of explainability and demonstrates high performance across four distinct criteria.

Finally, recent work by Mayda and Uğurlu [12] demonstrated the effectiveness of Transformers for Turkish reviews. Our study confirms their findings on Japanese data while contributing a novel multi-label framework for explainability, which is often missing in purely performance-focused studies.

**Table 1**
Chronological Summary of Key Studies in Review Usefulness Prediction.

| Reference | Year | Language | Dataset Type | Task Type |
|---|---|---|---|---|
| Kim et al. [1] | 2006 | English | Amazon | Regression |
| Mudambi & Schuff [2] | 2010 | English | Amazon | Regression |
| Zhang & Tran [3] | 2010 | N/A | N/A | Ranking |
| Pan & Zhang [4] | 2011 | English | Amazon | Regression |
| Sasaki et al. [5] | 2014 | Japanese | N/A | Classification |
| Qu et al. [6] | 2018 | English | Amazon | Classification |
| Zhang & Lin [7] | 2018 | English, German | Amazon | Classification |
| Sun et al. [8] | 2019 | Chinese | Dianping | Classification |
| Saumya et al. [9] | 2020 | English | Amazon | Regression |
| Malik [10] | 2020 | English | Amazon | Regression |
| Soda et al. [11] | 2021 | Japanese | N/A | Classification |
| Mayda & Uğurlu [12] | 2024 | Turkish | E-commerce | Classification |
| **Our Study** | – | Japanese | Rakuten, Amazon | Classification (Binary and Multi-label) |

### 2.1. Summary and Contributions

Prior research on review helpfulness evolved from traditional machine learning with engineered features to end-to-end deep learning models, typically for regression or binary classification tasks. Our work builds on this foundation by using Transformers not just for prediction, but to create an explainable system. We move beyond a single predictive score to identify the specific, human-understandable characteristics that make a review valuable.

The main contributions of this research are:

1. **A novel information score ('lds')** based on lexical density and text length, designed for efficiently sampling information-rich reviews from large datasets.
2. **A refined set of six objective criteria** of review helpfulness for Japanese.
3. **An explainable multi-label framework** based on those six criteria that identifies why a review is useful.
4. **A new, publicly available dataset** of over a thousand fully annotated Japanese reviews annotated with both binary usefulness and multi-label feature labels for explainable review analysis.

## 3. Methodology

We designed our approach to create an explainable, data-driven system for assessing review helpfulness. This involved three main stages: first, establishing a clear and comprehensive framework for what constitutes a useful review; second, developing a novel sampling method to efficiently curate high-quality data; and third, creating a high quality manually annotated dataset to train and evaluate our models. This section details each of these stages.

### 3.1. A Framework for Review Helpfulness

Building on prior research into the characteristics of helpful reviews [5, 11], we define a useful review as one that satisfies a set of specific, identifiable criteria. These criteria were developed by synthesizing

the findings of previous studies. Based on a preliminary analysis into review usefulness [13, 14], we established that the following six conditions characteristic for a review to be considered helpful. However, instead of assuming that all of the condition must be met, we analyzed which exact combinations of the following features make a review feel helpful. See Section 3.3.1 for details of this analysis.

Specifically, for a review to be considered helpful it must contain the following features in various combinations.

- A1 **A basis is provided for the evaluative expression.** The review explains why a product, or its aspect, was rated positively or negatively, going beyond simple evaluative statements.
- A2 **There are multiple mentions of the review target.** The review remains focused on the product itself, and does not digress about other products.
- A3 **The star rating and the evaluation in the review body are consistent.** The sentiment of the text aligns with the given star rating, ensuring the review is coherent.
- A4 **The main part of the review has a sufficient amount of information.** The review offers enough detail for a reader to make an informed decision.
- A5 **The review title contains the polarity and its target.** The title functions as an effective summary of the review's core message.
- A6 **It is possible to infer whether the reviewer actually used the product.** The text contains descriptions of first-hand experience with the product.

Egawa's work on text structure suggests that the title often serves as a reliable summary of a review's main point [15]. We hypothesize that a well-written, and informative title is a strong indicator of a well-written, informative review body. On the other hand, we deliberately excluded criteria such as "comparison with other products" and "readability", which were used in some prior works [5, 11]. This decision was based on findings that "comparison" sentences are often too rare to be a reliable feature, and "readability" is too subjective and lacks a concrete, consistently applicable definition in the existing literature. Instead, to also cover the lexical richness of the review, we proposed and 'lds' score for pre-filtering of low-quality reviews, as described in the following sections, which we used in the creation of the dataset used in this research.

## 3.2. Data Curation and LDS Sampling

To create a high-quality dataset for annotation without manually reading through millions of reviews, we first developed a method to sample information-rich content from large, unlabeled corpora. This was necessary because a purely random sample would likely be dominated by short, uninformative, and ultimately not useful reviews. For this we used data from the Amazon Review Dataset[1] and the Rakuten Ichiba Review Dataset [16].

We based our sampling method on a novel information score that adapts the concept of lexical density with additional penalization for too short texts.

Lexical density [17, 18, 19] measures the vocabulary richness of a text and is defined as the ratio of unique words ($wt$) to the total word count ($wc$), as shown in Equation 1. A higher score indicates a more diverse vocabulary.

$$ld = \frac{wt}{wc} \tag{1}$$

However, lexical density alone is biased towards shorter texts, which naturally have fewer repeated words. To counteract this, we introduce a penalty for short sentences. We first normalize the word count ($wc$) using a logarithmic scale to reduce the impact of extreme outliers (Equation 2).

$$wc_l = \log_2(wc) \tag{2}$$

[1] http://web.archive.org/web/20201127140619/https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_JP_v1_00.tsv.gz, accessed on 2025-10-17.

We then create a short sentence coefficient ($ssc$) by scaling this value by the maximum log-transformed word count in the corpus (Equation 3). This coefficient approaches 1 for the longest reviews and is smaller for shorter ones.

$$ssc = \frac{wc_l}{wc_{l_{max}}} \tag{3}$$

Finally, we define our proposed review information score (‘lds‘) by multiplying the lexical density by this short sentence coefficient (Equation 4). This score balances vocabulary richness with review length, favoring texts that are both lexically diverse and sufficiently long.

$$lds = ld \times ssc \tag{4}$$

Next, we applied the newly defined information score to create an initial dataset of review candidates. We used data from the one month of Rakuten Dataset (January 2019), from which we randomly extracted 10,000 review samples with 2 or more thumbs-up, indicating that two or more users found these reviews helpful, and another 10,000 samples with 0 thumbs-up (no users fund the review helpful), and used this data for validation to the proposed information score and extraction of training data for machine learning experiments.

The reason for setting the threshold to "2 or more thumbs-up" was that there was an overwhelming number of reviews with "1 or more thumbs-up" (38,734), which would account for most of the reviews with thumbs-up (54,861), making the random extraction biased. Reviews with "2 or more thumbs-up" accounted for 16,127 reviews with thumbs-up counts from 2 to 159, which assured a balanced and varied source for extraction.

Finally, each review sample from the initial dataset was assigned the ’lds’ information score. For our main experiment, we applied this score to the verification dataset and selected the top 1,000 unique reviews for manual annotation. This practical application of an advanced informativeness metric saves significant time and resources in the annotation process by pre-filtering for content that is more likely to be useful.

### 3.3. Annotation Process and Dataset Creation

Using the ‘lds‘ score, we selected the top 1,000 reviews from our verification dataset for manual annotation. The goal was to create a gold-standard dataset for training and evaluating automatic classification models for subjective helpfulness and its representative objective categories.

The annotation was conducted by 20 annotators (3 males, 17 females, all in their 20s to 40s) recruited through the CrowdWorks crowdsourcing platform[2]. Each annotator was provided with a detailed set of guidelines, instructing them to perform two tasks for each review sample:

1. Assign a binary label ("helpful" or "not helpful") based on their subjective judgment of whether the review would be helpful in a purchasing decision.
2. Assign a multi-label annotation by selecting all applicable criteria from our six-point framework (described in Section 3.1).

To assure high quality of the dataset, each review sample was annotated by two different people. The average of inter-annotator agreement (kappa value) for all pairs of annotators was 0.571 with standard deviation of 0.069, which indicates a moderate and stable agreement. Considering that the agreement was calculated from the whole annotation task, namely, both the subjective helpfulness and the set of objective features, meaning that the task was highly sophisticated, this level of agreement can be considered sufficiently high, suggesting high quality of annotations by all participants. Finally, the disagreements were resolved through a discussion with a well trained super-annotator.

[2] https://crowdworks.jp/

**Table 2**
Top 5 feature combinations in reviews judged as "Helpful" (left) and "Not Helpful" (right).

| Top 5 combinations in "Helpful". | | | Top 5 combinations in "Not helpful". | | |
|---|---|---|---|---|---|
| Feature Combination | Helpful | Not Helpful | Feature Combination | Not Helpful | Helpful |
| 1+2+3+4+6 | 263 | 0 | 1+2+3+6 | 79 | 197 |
| 1+2+3+6 | 197 | 79 | No matching features | 71 | 0 |
| 2+3+6 | 31 | 52 | 2+3+6 | 52 | 31 |
| 1+2+6 | 14 | 8 | 3 | 46 | 0 |
| 1+2+3 | 12 | 14 | 1+3+6 | 24 | 8 |

**Table 3**
Final composition of the annotated dataset.

| | Helpful | Not helpful | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|---|---|
| Training: Rakuten (1,000) | 574 | 426 | 683 | 779 | 821 | 295 | 28 | 782 |
| Test (in-domain): Rakuten (200) | 121 | 79 | 158 | 169 | 173 | 96 | 27 | 150 |
| Test (cross-domain): Amazon (192) | 96 | 96 | 107 | 117 | 108 | 45 | 80 | 164 |

### 3.3.1. Analysis of the Annotated Dataset

The final annotated dataset consists of **574** "useful" and **426** "not useful" reviews. An analysis of the co-occurrence of our defined features reveals important patterns. Table 2 shows the top five most frequent combinations of features for useful and not useful reviews, respectively.

These tables reveal a clear and compelling pattern. The combination ‘1+2+3+4+6‘, which includes providing a basis for evaluation, mentioning the target, being consistent, having sufficient information, and describing actual use, was found in 263 reviews, and every single one of them was labeled "useful." This combination represents a ’gold standard’ for a high-quality review. In stark contrast, reviews that matched none of our features or only matched the "consistency" feature (Label 3) were overwhelmingly labeled "not useful." This analysis validates that our framework successfully captures the core elements that align with human perceptions of helpfulness and provides a strong empirical basis for our multi-label classification task.

### 3.3.2. Final Dataset Composition and Limitations

The final dataset is composed of 1,200 annotated reviews from the Rakuten dataset (1,000 for training, 200 for evaluation), along with a separate 192-review set from Amazon used for cross-domain testing. Each entry includes the review text and two sets of labels: a binary "useful/not useful" label and six binary labels for our defined features. The final composition of the dataset was represented in Figure 3.

A key limitation, however, is the sampling bias introduced by our ‘lds‘ scoring method. The dataset is intentionally enriched with reviews that are longer and more lexically diverse. Although this pre-filtering phase can function as a component of the review helpfulness estimation method as a whole, in practice models trained on this data may not generalize perfectly to a random, unfiltered sample of all reviews. This trade-off between sample quality and representativeness should be considered when interpreting the model’s performance.

## 4. Experiments and Results

To evaluate our framework, we conducted two main experiments. In the first experiment we tested the overall concept of helpfulness by training a binary classifier to distinguish "helpful" from "not helpful" reviews. In the second, more fine-grained experiment we evaluated our primary contribution, namely,

an explainable multi-label model designed to identify the specific characteristics of a useful review based on our six-point framework.

## 4.1. Preliminary Investigation

First, we prepared review sentences and had two annotators verify if they met our definitions, using a small subset of 192 reviews from the Amazon review dataset[3]. The annotators also annotated whether the reviews were helpful or not based on these definitions.

The result was 96 helpful and 96 not helpful reviews. A1 to A6 in Table 4 correspond to the definition numbers in our study.

**Table 4**
Results of Preliminary Annotation on 192 Amazon Reviews

| | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| Helpful | 85 | 92 | 80 | 42 | 51 | 96 |
| Not Helpful | 22 | 25 | 28 | 3 | 29 | 68 |

From the results in Table 4, for definition "A6. It is possible to infer whether the reviewer actually used the product," 100% of reviews judged as helpful met this criterion, but many not helpful reviews also met this criterion. For "A1. A basis is provided for the evaluative expression," "A2. There are many mentions of the review target," and "A3. The star rating and the evaluation in the review body are consistent," a high percentage of useful reviews met these criteria, while only few not useful reviews did. The item "A5. The review title contains the polarity and its target" was met by more than half of the helpful reviews, confirming a certain degree of effectiveness. For "4. The review sentence has a sufficient amount of information," the number of matches was generally low in this annotation. For this preliminary study we set the threshold to 6-7 sentences or over 200 characters, which could influence this result. However, the difference between helpful and not helpful reviews was the strongest for this item, suggesting that the concept of pre-filtering, which we proposed in section 3.2, while not solving the review helpfulness problem alone due to having a high degree of misses, could function as a powerful pre-filtering tool when a large number of reviews is available.

## 4.2. Experimental Setup

Both experiments were built on state-of-the-art Transformer architectures. We selected a range of publicly available Japanese language models from the HuggingFace platform[4], as listed in the list below, to ensure a comprehensive evaluation.

Model 1 izumi-lab/electra-small-japanese-discriminator
Model 2 ku-nlp/roberta-base-japanese-char-wwm
Model 3 hiroshi-matsuda-rit/bert-base-japanese-basic-char-v2
Model 4 tohoku-nlp/bert-base-japanese-char-v2

All models were fine-tuned using a consistent set of hyperparameters, with learning rates of 1e-4, 1e-5, and 2e-5. The primary evaluation metric was the F1-score, supplemented by accuracy, precision, and recall. For the multi-label task, we report these metrics for each individual label.

## 4.3. Experiment 1: Binary Classification of Subjectively Perceived Helpfulness

The first experiment was designed to assess whether a model could learn a generalizable concept of review helpfulness. We framed this as a binary classification task in a challenging cross-domain

[3]https://aws.amazon.com/jp/blogs/news/learning-to-rank-amazon-customer-reviews/
[4]https://huggingface.co/

**Table 5**
Binary classification results (5-run average) for the cross-domain evaluation (Train: Rakuten, Test: Amazon).

| | LR: 1e-4 | | | | LR: 1e-5 | | | | LR: 2e-5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| Model 1 | 0.803 | 0.960 | 0.731 | 0.830 | 0.817 | 0.923 | 0.763 | 0.835 | 0.820 | 0.925 | 0.765 | 0.837 |
| Model 2 | 0.500 | 1.000 | 0.500 | 0.667 | 0.817 | 0.948 | 0.752 | 0.838 | 0.815 | 0.942 | 0.751 | 0.836 |
| Model 3 | 0.500 | 1.000 | 0.500 | 0.667 | 0.763 | 0.915 | 0.702 | 0.794 | 0.757 | 0.929 | 0.692 | 0.793 |
| Model 4 | 0.525 | 0.996 | 0.515 | 0.678 | 0.777 | 0.943 | 0.708 | 0.809 | 0.747 | 0.938 | 0.679 | 0.787 |

setting. The model was fine-tuned on the 1,000 reviews from our annotated Rakuten training dataset and evaluated on the 192-review Amazon test dataset. This setup tests the model's ability to transfer its learned knowledge from one e-commerce platform to another.

#### 4.3.1. Results

Table 5 presents the 5-run average performance for each model across the different learning rates. The results show that with an optimal learning rate (1e-5 or 2e-5), most models perform well, achieving F1-scores around or above 0.80. The best performance was achieved by the 'ku-nlp/roberta-base-japanese-char-wwm' model with a learning rate of 1e-5, reaching an F1-score of 0.838. A higher learning rate of 1e-4 proved unsuitable for most base-sized models, leading to classification collapse. The strong cross-domain performance indicates that the models successfully learned robust, platform-independent features of a useful review.

### 4.4. Experiment 2: Multi-label Classification of Objective Helpfulness Features

The second experiment was central to our goal of creating an explainable system for extracting helpful reviews. We framed this as an in-domain, multi-label classification task to train a model that can automatically detect which of our six defined helpfulness criteria are present in a given review. For this task, we used our full annotated Rakuten dataset of 1,200 reviews, split into a 1,000-review training set and a 200-review test set. The input to the model was a single text sequence created by concatenating the product name, review rating, title, and body.

#### 4.4.1. Results

Table 6 shows the performance of the top three models on this multi-label task. The results are exceptionally strong for four of the six labels. For Labels 1, 2, 3, and 6, the models consistently achieve F1-scores near or above 0.90, indicating that these criteria correspond to clear and detectable linguistic patterns.

The performance for Label 4 ("sufficient amount of information") was lower, with a max F1-score of 0.725. This is likely due to the inherent subjectivity of this criterion and a moderate data imbalance (a 3:7 ratio of positive to negative examples). Performance for Label 5 ("title contains polarity and target") was 0.00 across all models. After a deeper analysis, we found out that this was a direct result of our experimental setup. Specifically, there were only 28 of such cases within the whole 1,000-sample dataset, which was insufficient for the model to learn properly this criterion. Secondly, by concatenating all text into a single sequence, the model was prevented from distinguishing the title from the body. This highlights a clear area for future improvement in the input representation.

### 4.5. Case Study: Model Application at Scale

To demonstrate the practical application of our models, we conducted an exploratory case study on a large, unlabeled dataset of Rakuten reviews from January 2017. We first applied our binary helpfulness classifier and then used our multi-label model to analyze the features of the reviews within each class.

**Table 6**
Multi-label classification results for the top three models on the in-domain Rakuten test set.

| | Model 4 | | | Model 3 | | | Model 2 | | | Model 1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LR = 1e-5, Test loss = 0.419 | | | LR = 2e-5, Test loss = 0.425 | | | LR = 1e-5, Test loss = 0.436 | | | LR = 1e-4, Test Loss = 0.471 | | |
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| Label 1 | 0.887 | 0.892 | 0.890 | 0.870 | 0.930 | 0.899 | 0.895 | 0.861 | 0.877 | 0.893 | 0.842 | 0.866 |
| Label 2 | 0.917 | 0.917 | 0.917 | 0.910 | 0.953 | 0.931 | 0.925 | 0.953 | 0.939 | 0.919 | 0.876 | 0.897 |
| Label 3 | 0.881 | 0.988 | 0.932 | 0.891 | 0.994 | 0.940 | 0.907 | 0.960 | 0.933 | 0.894 | 0.931 | 0.912 |
| Label 4 | 0.827 | 0.646 | 0.725 | 0.893 | 0.521 | 0.658 | 0.823 | 0.531 | 0.646 | 0.907 | 0.406 | 0.561 |
| Label 5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Label 6 | 0.823 | 0.933 | 0.875 | 0.846 | 0.953 | 0.897 | 0.831 | 0.980 | 0.899 | 0.851 | 0.913 | 0.881 |
| Average | 0.723 | 0.729 | 0.723 | 0.735 | 0.725 | 0.721 | 0.730 | 0.714 | 0.716 | 0.744 | 0.661 | 0.686 |
| Average (w/o Label 5) | 0.867 | 0.875 | 0.868 | 0.882 | 0.870 | 0.865 | 0.876 | 0.857 | 0.859 | 0.893 | 0.794 | 0.823 |

**Table 7**
Top 3 predicted feature combinations for helpful vs. not helpful reviews on Rakuten Dataset.

| Helpful Feature Combinations | Helpful | Not Helpful | Not Helpful Feature Combinations | Not Helpful | Helpful |
|---|---|---|---|---|---|
| 1.+2.+3.+4.+6. | 114,096 | 10,386 | 1.+2.+3.+6. | 283,930 | 93,219 |
| 1.+2.+3.+6. | 93,219 | 283,930 | 3. | 232,898 | 2,063 |
| 2.+3.+6. | 28,66 | 77,875 | No matching items | 106,241 | 59 |

This is not a formal validation of accuracy, but a qualitative analysis to see if the patterns learned from our small annotated dataset hold true at a much larger scale.

Our binary model classified **213,142** reviews as "helpful" and **735,393** as "not helpful." We then analyzed the feature combinations predicted by the multi-label model for each class. The results, shown in Table 7, reveal patterns that are remarkably consistent with our findings from the manual annotation (Table 2). The combination ‘1+2+3+4+6‘ is overwhelmingly associated with "helpful" reviews, while reviews with "no matching items" or only "Label 3" are strongly associated with "not helpful" reviews. This qualitative consistency suggests that our models have learned meaningful and generalizable patterns of review quality that align with human judgments, even when applied to a massive, unlabeled dataset.

# 5. Discussion

## 5.1. Interpretation of Results and Comparison with Prior Work

Our Transformer models achieved an F1-score of over 0.83 in a cross-domain evaluation, confirming the effectiveness of deep learning for this task as shown in recent work [12]. The successful generalization from the Rakuten to the Amazon dataset suggests the model learned robust, platform-independent linguistic features of review usefulness.

The primary contribution of our work is its explainable multi-label framework. High F1-scores (over 0.9) for criteria such as "provides a basis for evaluation" (Label 1) and "content indicates actual use" (Label 6) demonstrate that core components of a quality review are linguistically detectable. This provides a data-driven method for achieving the explainability goals of earlier work [11]. The lower performance for "sufficient amount of information" (Label 4) likely stems from the inherent subjectivity of this criterion and data imbalance, while the failure on "title contains evaluation" (Label 5) was a direct result of input representation, highlighting an area for future improvement.

A key finding is that information quality, not just quantity, is pivotal. The presence of "sufficient information" (Label 4) was the crucial differentiator for usefulness, offering a more nuanced insight than the simple "review length" heuristic used in previous studies [2, 3]. Our ‘lds‘ score provides a

better proxy for this qualitative informativeness than word count alone, a concept explored also in prior work [8]. Furthermore, by focusing only on review content, our approach offers broader applicability than methods that depend on reviewer metadata [4, 10].

### 5.2. Methodological Considerations and Limitations

Our study has several limitations that provide clear avenues for future research. Firstly, our dataset was curated using the ‘lds‘ score, which introduces a sampling bias towards information-dense reviews. Consequently, the model’s performance may not generalize to unfiltered data rich in short, simple texts, although the ‘lds‘ score itself could function as an effective pre-filter in a practical system.

Secondly, the binary classification was a cross-domain evaluation (training on Rakuten, testing on Amazon). While the strong results suggest generalization, this setup makes it difficult to separate model performance from the effects of domain shift. An in-domain evaluation is needed to establish a clearer performance baseline.

Finally, our modeling approach has technical limitations. The model’s failure to detect title features (Label 5) was caused by concatenating all text fields. In the future we plan to use structured inputs to solve this issue. Additionally, our content-only focus, while making the model more universally applicable, ignores reviewer metadata, which other studies have shown to be a strong predictor of helpfulness [4, 10]. This could be implemented to further improve the performance.

### 5.3. Theoretical and Ethical Implications

From a theoretical perspective, our work frames "helpfulness" as a multi-faceted construct rather than a monolithic score, providing a data-driven validation for multi-perspective frameworks like that of Soda et al. [11]. Furthermore, our findings suggest that information density and lexical diversity, as captured by our ‘lds‘ score, are more precise indicators of a review’s quality than the simple review length heuristic used in many previous studies [2, 3], shifting the focus from quantitative to qualitative content metrics.

Ethically, while the intended application is to empower consumers by reducing information overload, deploying such a system always carries significant risks. These include bias amplification, where the model could systematically favor certain writing styles and marginalize others; the potential for malicious actors to game the system by crafting fake reviews optimized to our criteria; and algorithmic gatekeeping, where the system might suppress unconventionally written but helpful reviews. Therefore, transparent communication of the criteria for "helpfulness" is essential to ensure such systems promote fair discourse rather than implicitly censoring it.

## 6. Conclusions and Future Work

This study presented and validated an end-to-end framework for both predicting the subjective helpfulness of online product reviews and for providing and objective explainable basis of this prediction. We introduced a multi-label classification approach based on a refined set of six usefulness criteria. To facilitate this, we developed a novel ‘lds‘ score for sampling information-rich data and created a new, manually annotated dataset of over a thousand Japanese reviews.

Our experiments demonstrate the effectiveness of this framework. A Transformer-based binary classifier achieved a high F1-score of 0.838 in a challenging cross-domain evaluation, indicating that it learned generalizable features of usefulness. More importantly, our multi-label model successfully identified key criteria with F1-scores exceeding 0.9, confirming that these aspects are linguistically distinct and can be reliably detected. A key finding from our analysis is that while many reviews contain basic structural elements, the presence of "sufficient information" was the critical feature that distinguished helpful from non-helpful reviews, which is a more nuanced insight than the simple correlation with review length.

In conclusion, this research validates a multi-faceted, explainable approach as a powerful method for assessing review quality. By moving beyond a single predictive score, our framework provides a foundation for developing more transparent and effective systems to help consumers navigate the vast landscape of online feedback.

Future work will focus on two main areas, namely, dataset expansion and methodological refinement. A primary priority is to create a larger and more representative dataset that is not limited by our 'lds' sampling bias. On the methodological side, we also plan to address the failure in classifying title features (Label 5) by exploring structured inputs that distinguish the title from the body. Furthermore, a formal in-domain evaluation will be conducted to establish a clear performance baseline, complementing our current cross-domain results. Once these improvements are implemented, we will leverage the refined models to semi-automatically expand the annotated dataset, further scaling our research.

## Declaration on Generative AI

During the preparation of this work, the authors used Gemini 2.5 Pro in order to correct grammar and spelling.

## References


[1] S.-M. Kim, P. Pantel, T. Chklovski, M. Pennacchiotti, Automatically assessing review helpfulness, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06, Association for Computational Linguistics, 2006, p. 423. URL: http://dx.doi.org/10.3115/1610075.1610135. doi:10.3115/1610075.1610135.

[2] S. M. Mudambi, D. Schuff, What makes a helpful online review? a study of customer reviews on amazon.com, in: MIS quarterly, 2010, pp. 185–200.

[3] R. Zhang, T. Tran, Helpful or unhelpful: A linear approach for ranking product, Journal of Electronic Commerce Research 11 (2010) 220–230.

[4] Y. Pan, J. Q. Zhang, Born unequal: a study of the helpfulness of user-generated product reviews, Journal of retailing 87 (2011) 598–612.

[5] Y. Sasaki, Y. Seki, Shōhin rebyū wo taishō to shita yūyō-sei no teigi to hanbetsu [definition and discrimination of usefulness for product reviews], in: DEIM Forum B5-1, 2014.

[6] X. Qu, X. Li, J. R. Rose, Review helpfulness assessment based on convolutional neural network, arXiv preprint arXiv:1808.09016 (2018).

[7] Y. Zhang, Z. Lin, Predicting the helpfulness of online product reviews: A multilingual approach, Electronic Commerce Research and Applications 27 (2018) 1–10.

[8] X. Sun, M. Han, J. Feng, Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products, Decision Support Systems 124 (2019) 113099.

[9] S. Saumya, J. P. Singh, Y. K. Dwivedi, Predicting the helpfulness score of online reviews using convolutional neural network, Soft Computing 24 (2020) 10989–11005.

[10] M. S. I. Malik, Predicting users' review helpfulness: the role of significant review and reviewer characteristics, Soft Computing 24 (2020) 13913–13928.

[11] H. Soda, Shōhin rebyū no fukusū no kanten kara no yūyō-sei no hyōka [Evaluation of the usefulness of product reviews from multiple perspectives], Master's thesis, School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 2021.

[12] İ. Mayda, Y. Uğurlu, E-ticaret Sitelerindeki Türkçe Müşteri Yorumlarının Faydalılık Tahmini Predicting the Usefulness of Turkish Consumer Reviews on E-commerce Websites, in: 2024 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2024, pp. 1–5. URL: http://dx.doi.org/10.1109/ASYU62119.2024.10757106. doi:10.1109/asyu62119.2024.10757106.

[13] Y. Nakajima, M. Ptaszynski, F. Masui, Rebyū no yūyō-sei ni kansuru chōsa oyobi rebyū-bun no jōhō-ryō sukoa no teian [an investigation of review helpfulness and a proposal for an information

score for review sentences], in: LAU Technical Reports (Summer 2023), Language Acquisition and Understanding, Sapporo, Japan, 2023, pp. 21–30.
[14] Y. Nakajima, M. Ptaszynski, F. Masui, Rebyū no yūyō-sei ni okeru tokuchō bunseki oyobi transformers o mochiita rebyū no yūyō-sei hantei [feature analysis of review helpfulness and helpfulness classification using transformers], in: LAU Technical Reports (Summer 2024), Language Acquisition and Understanding, Kushiro, Japan, 2024, pp. 55–63.
[15] Y. Egawa, S. Konno, Bunshō kōsei o kōryo shita rebyū p/n bunrui shuhō no teian [a proposal of a p/n classification method for reviews considering document structure], in: IEICE Conferences Archives, The Institute of Electronics, Information and Communication Engineers, 2012.
[16] Rakuten Group, Inc., Rakuten dētasetto [rakuten dataset], https://doi.org/10.32130/idr.2.0, 2014.
[17] J. Ure, Lexical density and register differentiation, Applications of linguistics 23 (1971) 443–452.
[18] J. Eronen, M. Ptaszynski, F. Masui, A. Smywiński-Pohl, G. Leliwa, M. Wroczynski, Improving classifier training efficiency for automatic cyberbullying detection with feature density, Information Processing & Management 58 (2021) 102616.
[19] M. Ptaszynski, A. Yagahara, Terminology extraction device, terminology extraction method and program, Patent no.: 7557770 (2024-09-19).