修士論文

商品への言及に基づくレビューの有用性判定

漆原和輝

主指導教員　白井清昭

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和８年３月

Abstract

With the widespread of e-commerce (EC) platforms, product reviews have become an important source of information that strongly influences consumers' purchasing decisions. However, EC sites receive an enormous number of reviews every day, and many reviews do not contain concrete descriptions of product quality or usability. For example, some reviews only include complaints about packaging and delivery service, such as "the delivery was slow," or only describe personal circumstances or shipping experiences, such as "I bought this for my daughter." While such reviews may be meaningful as information about purchasing background or delivery experiences, they do not provide any information about the product itself.

In previous studies on identifying helpfulness of reviews, the dominant approaches estimated helpfulness using surface-level features such as review length, lexical diversity, and sentiment polarity, as well as external indicators such as the number of "helpful" votes. However, these studies did not explicitly evaluate the helpfulness of reviews from the perspective of whether the review mentions the product itself.

This study aims to classify the helpfulness of reviews by focusing on two key aspects: whether "a review mentions the product itself," and whether "a review contains a subjective evaluation of the product." Specifically, this thesis addresses a three-class classification task. The first class is "NotMentioned," which includes reviews that describe only delivery conditions or purchase motivations without mentioning the product itself. The second class is "Mentioned/NotEvaluated," which includes reviews that refer to product features or specifications but do not express a positive or negative evaluation. The third class is "Mentioned/Evaluated," which includes reviews that clearly evaluate product quality or usability.

A distinctive characteristic of the proposed approach is that it uses not only the review text itself but also auxiliary product information, namely the product description and the product name, as model inputs. This allows the model to consider the contextual relationship between the review and the product. In addition, because large-scale manual annotation for helpfulness classification is costly and difficult, pseudo-labeled training data is automatically constructed using a rule-based method and a large language model (LLM).

Two methods for constructing pseudo-labeled datasets are proposed. The first is a rule-based method. Important terms (keywords) are extracted from a corpus of product descriptions by computing TF-IDF scores, and then a review is classified whether it contains mention to a product by detecting these terms in a review. On the other hand, the presence of evaluation for a product in a review is identified by checking whether evaluation words in an evaluation lexicon appear in the review. Based on these results, each review is assigned to one of the

aforementioned three classes. The second method uses an LLM. By providing the review and product-related information in a prompt, the LLM determines which of the three classes the review belongs to. Preliminary experiments showed that the LLM-based method achieved higher labeling accuracy than the rule-based method. Therefore, pseudo-labeled data generated by the LLM were used as training data in the final experiments.

To build a classification model for helpfulness classification, pre-trained language models are fine-tuned. Two model architectures are developed: a one-stage approach that directly classifies reviews into three classes, and a two-stage approach that first determines whether a product is mentioned and then determines whether an evaluation is present. Preliminary experiments revealed that the two-stage approach suffered from error propagation, as misclassifications in the first stage could not be corrected in the second stage, resulting in lower accuracy. In contrast, the one-stage approach achieved more stable performance by learning features across all three classes simultaneously. Accordingly, the one-stage approach is applied in the succeeding experiments. In addition, two pre-trained language models are compared as the base model of the helpfulness classifier: BERT and ModernBERT. ModernBERT is an extension of the standard BERT model, which is better suited for handling longer contexts. It was found that ModernBERT achieved higher overall classification performance by preliminary experiments.

Several experiments were conducted to evaluate the effectiveness of the proposed method using a dataset from Rakuten Ichiba. Pseudo labels were assigned to create training data, while test data were constructed by manually annotating sampled reviews. Accuracy and macro F1 score were used as evaluation metrics, and training and inference were repeated three times to compute average scores. The experimental results showed that our proposed method achieved an accuracy of 74.2% and a macro F1 score of 69.5% at maximum. Next, several combinations of a review, product name, and product description are fed into the helpfulness classifier and the influence of the difference of these inputs was investigated. The results indicated that adding the product name to the model input improved classification performance under many conditions. This may be because product names often contain important features or keywords that help the model capture the relationship between the review and the product. In contrast, adding the product description did not always improve performance and sometimes degraded performance due to noise. Increasing the maximum input token length also did not necessarily lead to better performance.

Additionally, the proposed lightweight model based on ModernBERT was compared with zero-shot inference using GPT-4o. Although GPT-4o achieved a high accuracy of 79.6% when using "review + product description" as input, the pro-

posed method also achieved approximately 74% accuracy, resulting in a relatively small performance gap. Because LLMs such as GPT-4o require substantial computational resources due to their large number of parameters, their deployment in real-world services can be challenging. In contrast, the proposed method achieves comparable performance with significantly fewer parameters, making it more suitable for practical services that handle large volumes of reviews.

In summary, this study demonstrates that it is possible to construct a labeled dataset without manual annotation by leveraging LLMs, and to build a lightweight and accurate model for helpfulness classification based on product mentions and evaluations. By applying this model, EC platforms can efficiently present reviews that are more useful to consumers, thereby supporting users' purchasing decisions. Future work includes improving the quality of pseudo labels, further optimizing input design, and expanding evaluation datasets.

概要

EC サイトの普及に伴い，商品レビューは消費者の購買意思決定に大きな影響を与える情報源となっている．しかし，EC サイトには日々膨大な数のレビューが投稿されており，その中には商品そのものの品質や使用感に関する具体的な記述を含まないものも多数存在する．例えば，梱包の状態に対する不満，あるいは「娘のために購入しました」「発送が遅かったです」といった個人の事情や発送体験に関する情報のみを述べたレビューなどが該当する．これらは購買背景や配送体験に関する情報としては意味を持つものの，商品そのものの内容に関する情報ではない．従来のレビュー有用性判定に関する研究では，レビューの有用性を推定するために，テキストの長さ，語彙の多様性，感情極性といった表層的な特徴量や，「参考になった」ボタンの投票数などの外的指標を用いる手法が主流であった．しかし，これらの先行研究ではレビューが商品の内容について言及しているかという観点からレビューの有用性を判定するものではなかった．

本研究では，レビューが「商品の内容について言及しているか」，および「商品に対する主観的な評価を含んでいるか」という 2 つの観点に着目し，レビューの有用性を判定することを目的とする．具体的には，レビューを以下の 3 つのクラスに分類するタスクに取り組む．第一に，配送状況や購入動機のみを述べ，商品そのものに言及していない「言及なし」，第二に，商品の特徴や仕様について触れているが良し悪しの判断を含まない「言及あり・評価なし」，第三に，商品の品質や使用感について明確な評価を示している「言及あり・評価あり」である．本手法の特色は，レビューテキスト単体だけでなく，その対象となる商品の「商品説明文」や「商品名」を補助情報としてモデルに入力し，レビューと商品情報との文脈的な関連性を考慮して判定を行う点にある．また，有用性判定モデルの訓練データを構築するための大規模な人手アノテーションが困難であるという課題に対し，ルールベース手法および大規模言語モデル (LLM) を用いて擬似ラベル付きデータセットを自動構築する．

擬似ラベル付きデータセットを構築するために 2 つの手法を提案する．1 つ目はルールベース手法である．商品説明のコーパスから単語の TF-IDF を算出して重要語のリストを獲得し，この重要語がレビューに含まれるか否かで商品への言及の有無を判定する．一方，評価語辞書における評価語がレビューに含まれるか否かで評価の有無を判定する．これらの判定結果を踏まえてレビューに 3 つのクラスのいずれかを付与する．2 つ目は LLM を用いた手法である．レビューと商品説明など商品に関する情報をプロンプトとして与え，LLM にレビューが 3 つのクラスのいずれに該当するかを判定させる．予備実験の結果，LLM を用いた手法はルールベース手法に比べて高い精度でラベル付けができることを確認した．そのため，最終的に LLM によって擬似ラベルを付与したデータを訓練データとして使用した．

有用性判定を行う分類モデルの構築するために，事前学習済み言語モデルのファインチューニングを行った．モデル構造として，3 つのクラスを一度に分類する

「一段階方式」と，まず言及の有無を判定し，次に評価の有無を判定する「二段階方式」の 2 種類を比較検討した．予備実験の結果，二段階方式では第 1 段階での誤判定（言及があるにもかかわらず「なし」と判定される）を第 2 段階で訂正することはできないため，正解率が低いことが明らかになった．一方，一段階方式は 3 クラス間の特徴を同時に学習することで安定した性能を示した．そのため，本研究では一段階方式を採用した．また，事前学習済み言語モデルとして，標準的な BERT と，より長い文脈の処理に適した ModernBERT を比較したところ，ModernBERT の方が全体的に分類の性能が高かった．

提案手法の有効性を評価する実験を行った．実験には楽天市場のデータセットを使用し，擬似ラベルを付与して訓練データを作成した．また，楽天市場のデータセットからレビューをサンプリングし，人手でラベル付けを行ってテストデータを作成した．評価指標は正解率とマクロ F1 スコアとした．学習とテストを 3 回繰り返し，これらの評価指標の平均を求めた．実験の結果，正解率は 74.2%，F1 スコアは 69.5%という結果が得られた．さらに，入力情報の構成が判定の性能に与える影響を分析した．その結果,「商品名」をモデルの入力に追加することは多くの条件で分類性能の向上に寄与することが確認された．これは，商品名の中に商品の重要な特徴やキーワードが含まれていることが多く，モデルがレビューと商品との関連性を捉える助けになったためと考えられる．一方で,「商品説明文」を入力に加えることは必ずしも分類性能の向上につながらず，場合によっては不要な情報がノイズとなり性能を低下させる傾向が見られた．また，最大入力トークン長を増やしてより多くの情報を与えても，単純に性能が向上するわけではないことも明らかになった．

さらに，提案手法（ModernBERT を用いた軽量モデル）と，GPT-4o によるゼロショット推論の性能比較も行った．GPT-4o は「レビュー＋商品説明」を入力とした場合に 79.6%という高い正解率を示したが，提案手法も 74%程度の正解率を達成しており，その差はそれほど大きいものではなかった．GPT-4o のような LLM はパラメタ数が非常に多く，多大な計算機資源を必要とするため，実サービスへの導入が容易とは限らない．それに対し，提案手法はパラメタ数が大幅に少ないモデルで LLM と同等程度の性能を達成できており，大量のレビューを扱う実サービスに適用しやすいという利点を持つ．

以上をまとめると，本研究は LLM を活用して人手によるアノテーションなしでラベル付きデータセットを構築し，これを訓練データとして，商品への言及と評価の有無という観点からレビューの有用性を軽量かつ高精度に判定できるモデルを実現した．このモデルを応用することで，消費者が求めるレビューを効率的に提示し，EC サイトにおけるユーザの商品購入の意志決定を助けることが期待される．今後の課題としては，擬似ラベルの品質向上，入力設計のさらなる最適化，評価用データの拡充などが挙げられる．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

EC サイトにおける商品レビューは，消費者の購買意思決定に大きな影響を与える情報源である．Kuan らは購入前にレビューを確認する消費者は全体の 92%に達し，そのうち 89%が「レビューが購買決定に大きな影響を与える」と回答していることを報告している [1]．このように，レビューは購買行動を支える重要な役割を果たしている一方で，EC サイトには日々膨大な数のレビューが投稿されるため，利用者がすべてのレビューを閲覧し，有用な情報を取捨選択することは困難である．そのため，レビューの中から有用なものを自動的に判別し提示するレビュー有用性判定の研究が盛んに行われている．

従来のレビュー有用性判定の研究では，レビューの長さや語彙の多様性，感情極性といったレビュー自体の特徴量，あるいは「参考になった」投票数や返信数などの外的指標を用いて有用性を推定する手法が多く提案されてきた．しかし，これらの手法では，レビューが商品そのものについてどの程度具体的に述べているかという観点が必ずしも明示的に扱われていない．

例えば,「娘のために購入しました」や「発送が遅かったです」といったレビューは，購買背景や配送体験に関する情報としては意味を持つものの，商品そのものの内容に関する情報ではない．一方で，短い文でも,「サイズ感が小さめである」,「装着時のフィット感が良い」,「耐久性に不安がある」といった具体的な商品特徴を述べるレビューは，閲覧者にとって高い情報価値を持ち得る．このように，商品への言及の有無は，レビューの有用性を左右する重要な要因であると考えられる．

商品言及の観点を考慮せずに有用性判定を行うと，いくつかの問題が生じる．第一に，投票数などの外的指標は，投稿時期や閲覧数，プラットフォームの UI 設計の影響を受けやすく，レビュー内容の情報価値を必ずしも正確に反映しない可能性がある．第二に，レビューの長さや感情極性を考慮するだけでは，配送体験や購入動機に関する記述と，商品性能や使用感に関する具体的記述とを十分に区別できない場合がある．その結果，閲覧者が本当に知りたい「商品性能・品質・使用感」に関するレビューが埋もれ，レビュー閲覧の負担増大や誤った購買判断につながるおそれがある．したがって，レビュー有用性判定においては,「商品に言及しているかどうか」という観点を明示的に導入し，商品に関する具体的情報を含

むレビューを選別できる枠組みが求められる．

一方，上記の課題に取り組むにあたり，高品質な教師データの不足は大きな障壁となる．レビューの有用性判定は主観性を含むタスクであり，大規模な人手アノテーションには多大な時間的・金銭的コストがかかる．

1.2 目的

本研究の目的は，レビューが商品内容に言及しているかどうか，および商品に対する評価を含むかどうかの観点から，レビューの有用性を判定する手法を提案することである．具体的には，レビューを次の 3 クラスに分類する．

- **言及なし**：商品そのものについて言及しておらず，配送状況，購入動機，期待，あるいは投稿者自身の状況のみを述べたレビュー．
- **言及あり・評価なし**：商品に言及しているが，良し悪しなどの主観的評価を含まず，事実の記述や推測，第三者の反応の報告に留まるレビュー．
- **言及あり・評価あり**：商品の使用結果や品質について，満足・不満などの評価が明確に述べられているレビュー．

本タスクでは，レビュー自体の特徴だけでなく，レビュー対象商品の説明に含まれる内容と，レビュー内の記述との関連性を考慮することが重要である．そのため本研究では，レビューと商品説明を入力とする分類モデルを構築する．

また，分類モデルの教師あり学習のためのラベル付きデータの構築のコストが高いという問題に対し，大規模な人手ラベル付与を行う代わりに，ルールベース手法および大規模言語モデル (Large Language Model; LLM) を用いてレビューに擬似ラベルを付与し，訓練データを自動構築する枠組みを採用する．

さらに，ルールベースおよび LLM を用いた擬似ラベル付与手法の比較，後述する一段階方式・二段階方式といった分類手法の違い，事前学習済み言語モデルの選択が性能に与える影響などを実験的に検証することで，レビューの有用性判定の実用的なモデルを設計するための指針を示す．

1.3 本論文の構成

本論文の構成を以下に示す．2 章では，レビュー有用性判定に関する先行研究，ならびにレビューと商品情報の関連性に着目した研究を整理し，本研究の位置づけを述べる．3 章では，本研究で扱うタスク定義を示した上で，擬似ラベル付きデータセットの自動構築手法，および事前学習済み言語モデルを用いた有用性判定モデルの学習方法を提案する．4 章では，提案手法の評価実験について述べる．

擬似ラベル付与手法の比較，一段階方式と二段階方式の比較，ならびに有用性判定に用いる分類器の学習に用いる言語モデルの比較を通して提案手法の有効性を評価する．最後に 5 章で，本研究のまとめと今後の課題を述べる．

第2章　関連研究

本章では，本研究に関連する既存研究について整理する．2.1 節では，レビューの有用性判定に関する研究を概観し，これまでに用いられてきた特徴量やモデル構造について述べる．2.2 節では，レビューと商品情報の関連性に着目した研究を整理し「レビューが商品内容に言及しているかどうか」という観点の重要性を論じる．2.3 節では，LLM を用いたデータ構築およびアノテーションに関する研究について触れる．2.4 節では，長い入力を扱う言語モデルの発展について述べる．2.5 節では，これらの既存研究を踏まえ本研究の特色と位置づけを示す．

2.1　レビューの有用性判定

オンラインレビューの有用性判定は，EC サイトや口コミサイトにおいて，利用者にとって価値の高い情報を効率的に提示するための重要な研究課題である．従来の研究では，レビューがどの程度「役に立つか」を定量的に推定するため，さまざまな情報源や学習枠組みが検討されてきた．

McAuley と Leskovec はレビューのテキストと星評価の間には乖離が存在することを指摘し，テキストに隠されたユーザの評価を考慮することの重要性を説いている [2]．これは，星評価だけでは捉えきれない「内容の有用性」を解析する必要性を示している．

これまでの研究においては，レビューの有用性を決定づける要因として，テキストの長さ，語彙，感情極性といった内的指標（テキスト自体から得られる特徴量）に着目した分析が多く行われてきた．

Singh らは，Amazon のレビューデータを対象に勾配ブースティング決定木を用いた有用性予測モデルを構築した [3]．この研究ではレビューの極性や主観性といった感情的特徴に加え，テキストの読みやすさや情報量を測る指標としてのエントロピーに着目した．その結果，商品タイプを問わずこれらのテキスト特徴量が有用性の決定要因として極めて重要であることを示した．

高島らは，化粧品レビューサイト「@cosme」を対象に，ユーザからの「Like」投票数を有用性の指標と定義し，これを回帰モデルにより推定する手法を提案した [4]．彼らは，レビューの文字数や改行数といった構造的な特徴に加え，専門用語の数や Latent Dirichlet Allocation(LDA) を用いて推定された潜在的トピックなどの

意味的な特徴を素性として用いることで，単語の出現頻度（Bag-of-Words;BoW）のみを用いる場合と比較して，有用性推定性能が向上することを示した．

Hong らは，レビューの有用性に関する既存の多数の研究を対象にメタ分析を行い，レビューの長さやレビュワーの専門性などが有用性に一貫して正の影響を与える一方で，可読性などの要因は研究によって有効か否かの結論が異なることを報告した [5]．

一方で，内的指標だけでなく，投票数や返信，商品情報といった外的指標（テキスト以外の情報から得られる指標）を組み合わせることで，有用性を評価する手法も提案されてきた．

Qin らは，レビューの有用性を内的有用性と外的有用性の双方から包括的に評価する枠組みを提案した [6]．テキストの長さやセンチメントといった内的指標に加え，レビューに対する「返信」に含まれる否定的な情報の割合を新たな外的指標として導入した．後者は，従来の外的指標としてよく使われていた「役に立った」の投票数とは異なる新しい提案であった．

Fan らは，レビューテキストだけでなく，製品のメタデータ（タイトル，ブランド，カテゴリなど）とレビューの双方を考慮したニューラルネットワークモデル（PRH-Net）を提案し，製品情報を組み込むことが有用性予測の性能向上に寄与することを示した [7]．

近年自然言語処理の分野で優れた成果をあげている事前学習済み言語モデルをレビューの有用性判定に応用した研究も行われている．Xu らは，BERT [29] を用いてオンラインレビューの有用性スコアを予測するモデルを構築し，従来の機械学習手法と比較して高い性能を示した [8]．この研究は，BERT による文脈を考慮した表現学習がレビューの有用性判定に有効であることを示している．

また，有用性の判定だけでなく，レビューとして不適切な記述やノイズを排除するという観点からの研究も重要である．Jindal と Liu は，レビューを対象としたスパム検出タスクについて，ブランドへの言及がないものや内容が重複しているものをスパムとして定義し，その検出手法を提案した [9]．

以上より，既存のレビュー有用性判定研究では，文脈理解や表現学習の高度化が進んでいる一方で，レビューが商品内容に言及しているかどうかという観点は暗黙的に扱われることが多く，明示的な判定対象として十分に扱われていない．

2.2　レビューと商品情報の関連性

レビューの有用性をより本質的に評価するためには，レビュー単体だけでなく，レビューの対象である商品の情報との関係性を考慮することが重要である．本節では，レビューと商品説明，商品属性，画像などを統合的に扱う研究を概観する．

Hu と Liu は，レビューから「商品特徴」と「意見」を抽出し，特徴ごとの評判を要約する手法を提案した [10]．彼らは，アソシエーション分析を用いて頻出する

名詞句を商品特徴として特定し，WordNet の類義語・反義語関係を利用して形容詞の極性 (ポジティブ・ネガティブ) を自動判定するアルゴリズムを構築した．この研究は，商品の特徴 (属性) と評価を結びつける属性レベルの感情分析の先駆的な事例であり，レビューが商品のどの属性に言及しているかを構造化する試みの基礎となっている．

Ni らは，レビューを推薦システムの「説明資源」として利用する際，そのレビューが商品のどの側面について述べているかを特定することが重要であると述べている [11]．これは，有用なレビューには商品内容への言及が不可欠であることを示している．

Zhang らは，レビュー中の語と商品属性の対応関係を明示的にモデル化する Explicit Factor Model を提案し，説明可能な推薦を実現している [12]．

新井と佐藤は，レビュー中の意見文が「デザイン」や「携帯性」といった特定の評価視点に言及しているかどうかを関連語辞書を用いて判定する手法を提案した [13]．この研究は，レビューが商品の具体的な側面に言及している度合いを定量化する試みであるとみなせる．

松波らは，化粧品レビューサイトを対象とし，評価項目毎にユーザの評価点を推測する手法を提案した [14]．この研究では，総合的な星評価だけでは個々のユーザが重視する「うるおい」や「美白」といった具体的な観点を反映しきれないという課題に着目した．そして，レビューテキストから評価項目ごとの評価表現辞書を構築し，それを用いてレビューを観点別に自動採点する手法を提案した．この手法により，例えば「保湿力」を重視するユーザにはその項目が高く評価されているレビューを優先的に提示するなど，ユーザ個人のニーズに合わせた情報提供を実現している．

曽田と白井は，レビューが商品内容をどの程度説明しているかを「言及度」として定義し，商品カテゴリごとのキーワードに基づく TF-IDF を用いて言及度を推定する手法を提案した [15]．この研究は，レビューが商品にどれだけ深く言及しているかを定量的に捉えた点で意義がある．一方で，キーワードに基づく手法であるため，言い換え表現や文脈依存の記述を十分に扱えないという課題がある．

これらの先行研究では，レビューと商品情報の関連性はスコアやランキングとして扱われることが多く，レビューが商品内容に言及しているか否か，また，評価を含んでいるか否かを明示的なクラスとして整理する試みではない．

2.3 LLM を用いたデータ構築とアノテーション

深層学習モデルの学習には高品質かつ大量のラベル付きデータが不可欠であるが，人手によるアノテーションは時間的・金銭的コストが高いという課題がある．この課題に対し，自動的にアノテーションされた不完全なノイズを含むラベルを用いて学習を行う「弱教師あり学習」のアプローチが確立されつつある．Ratner

らが提案した「Snorkel」は，ルールベースやヒューリスティクスを用いて自動生成した擬似ラベルを確率的に統合し，学習データを構築するフレームワークであり，データ作成コストを大幅に削減できることを示した [16]．Meng らは，弱教師あり学習によって生成されたデータを用いても，ニューラルネットワークを用いたテキスト分類において実用的な精度が得られることを報告した [17].

近年では大規模言語モデル（LLM）をデータアノテーターとして活用する研究が急速に進展している．Wang らは，GPT-3 を用いてデータのラベリングを行うことで，人手によるコストを大幅に削減できる可能性を示した [18]．Ding らも同様に，GPT-3 が特定のタスクにおいて人間のアノテーターと同等以上の品質でデータを作成できることを示した [19].

Gilardi らは，ツイートやニュース記事を用いた関連性判定，スタンス検出，トピック分類といった複数のタスクにおいて，ChatGPT のゼロショットプロンプトを用いた擬似ラベル付与とクラウドソーシング（Amazon Mechanical Turk）によるラベル付与を比較検証した [20]．その結果，ChatGPT はクラウドワーカーよりも平均して高い正解率でラベルを付与できるだけでなく，アノテーター間の一致率においても訓練を受けた人間の作業者を上回る高い信頼性を示した．さらに，アノテーションコストはクラウドソーシングの 30 分の 1 以下に抑えられることを報告し，LLM が高品質かつ低コストなデータ構築の代替手段となり得ることを実証した.

また，既存データに対するラベリングだけでなく，LLM を用いて学習データそのものをゼロから構築する試みも行われている．Wang らは，少数の人手による指示データから LLM 自身の生成能力を利用して多様な指示と入出力のペアを大量に自動生成するフレームワーク「Self-Instruct」を提案した [21]．この手法は，人手によるデータ作成コストを最小限に抑えつつモデルの指示追従能力を向上させるためのデータ構築手法と位置付けられる.

これらの研究は，LLM が作成した「擬似ラベル」が，下流タスクのモデル学習において有効な教師データとなり得ることを示唆している．本研究では，これらの知見に基づき，GPT-4o を用いてレビューの疑似ラベルを自動生成し，分類モデルの訓練データとして利用する枠組みを採用する.

2.4 長い系列を扱う言語モデル

本研究では，レビューと商品説明を結合して有用性判定モデルに入力するため，長い入力系列を扱うことを考慮する必要がある．BERT [29] に代表される従来の Transformer モデルは，入力トークン数が最大 512 程度に制限されており，また Self-Attention 機構の計算量が系列長の二乗に比例するため，長文処理には適していなかった．この課題に対し，近年では計算効率を改善し，より長い文脈を扱えるモデルの開発が進んでいる.

Dao らは，FlashAttention と呼ばれる手法を提案し，GPU のメモリ転送を最適化することで Attention 機構の高速化と長文脈への対応を実現した [22].

計算効率の向上に加え，学習時よりも長い系列長に対してモデルが適応できるかという点も重要である．Press らは，正弦波位置埋め込みなどの従来手法が外挿性に乏しいことを指摘し，Attention の計算時にトークン間の距離に応じたペナルティを与えることで，学習時の系列長を超える入力に対しても性能を維持できる「ALiBi」を提案した [23].

一方で，モデルが扱えるコンテキスト長が増加しても，その中の情報を均質に利用できるとは限らない．Liu らは，長文脈を入力した際，モデルは入力の冒頭や末尾にある情報は正確に取得できるものの，中間に位置する情報の取得に関しては性能が著しく低下する「Lost in the Middle」という現象を報告した [24]．これは，単に入力可能なトークン数を増やすだけでなく，長い文脈内の情報を安定して処理するための評価や工夫が必要であることを示している．

ModernBERT [28] もまた，最新技術を取り入れ，従来の BERT よりも長いコンテキスト（最大 8192 トークン）を効率的に処理できるように設計されている．本研究では ModernBERT を有用性分類モデルとして用いる．これにより，商品説明のような長いテキストを含んだ入力に対しても，情報の欠落を防ぎつつ文脈を考慮した特徴抽出が可能となる．

2.5 本研究の特色

本研究の特色は，レビューの有用性を「商品内容への言及の有無」および「商品に対する評価の有無」という二つの観点から整理し，レビューを「言及なし」,「言及あり・評価なし」,「言及あり・評価あり」の 3 クラスに分類する枠組みを提案する点にある．従来のレビュー有用性判定研究では，レビューが商品内容をどの程度具体的に説明しているかという観点は，暗黙的に扱われることが多かった．本研究では，レビューと商品説明の対応関係に着目し，レビューが商品内容に関する情報を含んでいるかどうかを明示的に判定対象とすることで，有用性をより解釈しやすい形で定式化している．

なお，レビューの有用性は，わかりやすい文章であるか，競合商品との比較があるかなど，様々な観点から評価されうる．本研究では「レビューが商品に言及しているか」と「主観的な評価を含むか」という観点からレビューの有用性を判定するが，他の観点を無視するものではなく，他の観点による自動評価と相補的に用いることを想定している．

また，有用性判定モデルの学習データを構築のための大規模な人手アノテーションが困難であるという課題に対し，ルールベース手法および LLM を用いて擬似ラベル付き訓練データを自動構築する点も本研究の特徴である．

さらに，有用性判定モデルの構成として，3 クラスを直接分類する一段階方式と，言及判定と評価判定を分離する二段階方式を比較する．加えて，BERT と長

文入力を想定して設計されたModernBERTを比較することで，分類モデルとして使用する事前学習済み言語モデルの選択が性能に与える影響を実験的に検証する．

第3章　提案手法

3.1　タスクの定義

1.1 節で述べたように，本研究では，商品レビューと，当該レビューが対象とする商品の説明を入力とし，レビューが商品に言及してるか，レビューが商品に関する評価を含むかといった有用性の観点からレビューを分類するタスクを扱う．分類クラスの定義を再掲する．

- **言及なし**：商品そのものについて言及しておらず，配送状況や購入動機，自身や第三者の状態のみを述べたレビュー．
- **言及あり・評価なし**：商品について言及しているが，良し悪しなどの主観的評価を含まないレビュー．
- **言及あり・評価あり**：商品の使用結果や品質について，明確な評価を含むレビュー．

本タスクは，レビューと商品説明の文脈的な関連性を考慮する必要があり，単純なキーワードマッチングでは十分な性能を得ることが困難であると考えられる．

3.2　概要

提案手法の全体的な処理の流れを図 3.1 に示す．本研究で提案する手法は，レビュー集合に対して擬似ラベルを付与し，そのデータを用いて有用性判定モデルを学習する二段階構成からなる．

まず，商品説明とレビューの組に対し擬似的にラベルを付与したデータセットを構築する．次に，構築したデータセットを用いて，事前学習済み言語モデルをファインチューニングし，レビューの有用性を判定する分類器を学習する．

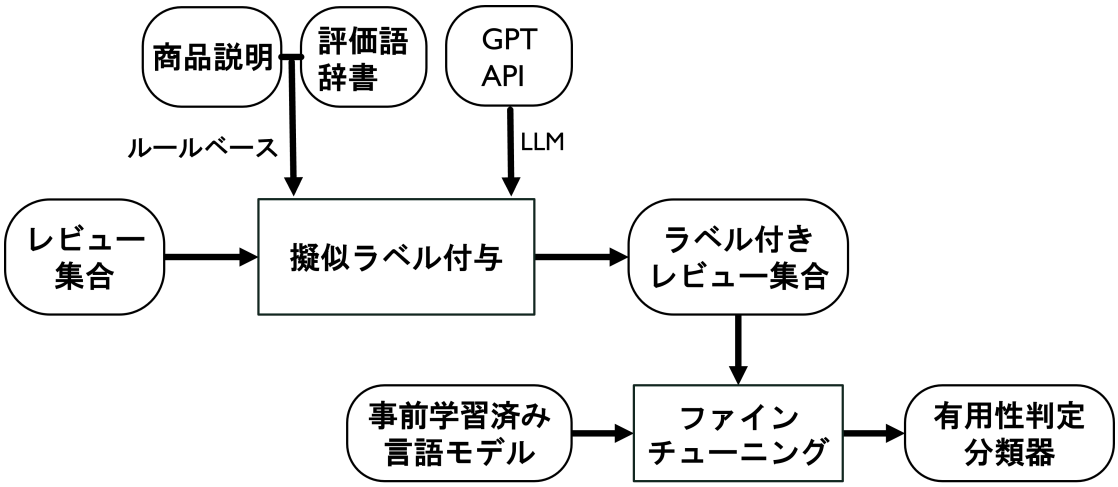


図 3.1: 提案手法の概要

3.3 訓練データの自動構築

本研究では，大規模なレビュー集合に対して人手でラベル付与を行うことが困難であるため，擬似ラベルを用いた訓練データの自動構築を行う．擬似ラベル付与には，ルールベースの手法および LLM を用いた手法の 2 つを提案する．

3.3.1 ルールによる疑似ラベル付与

レビューと商品説明の組に対し，ルールベースの擬似ラベル付与を行う．本手法は，(1) 商品への言及の有無の判定，(2) 評価の有無の判定を独立に行い，それらを組み合わせて 3 値ラベルを付与する．

ラベル付けルール　レビュー r と，そのレビューの対象商品の説明文書 d に対し，言及判定結果を $m(r,d) \in \{0,1\}$（0: 言及なし，1: 言及あり），評価判定結果を $e(r) \in \{0,1\}$（0: 評価なし，1: 評価あり）とする．最終ラベル $y(r,d)$ は次の規則で決定する．

$$y(r,d) = \begin{cases} \text{言及なし} & (m(r,d) = 0) \\ \text{言及あり・評価なし} & (m(r,d) = 1 \wedge e(r) = 0) \\ \text{言及あり・評価あり} & (m(r,d) = 1 \wedge e(r) = 1) \end{cases}$$

ここで，レビューが何らかの対象に対する評価を含んでいていも，商品への言及がない場合は「言及なし」とする点に注意していただきたい．

言及の有無の判定

言及の有無は，商品説明から抽出した「重要語」がレビューに出現するかどうかで判定する．手順を以下に示す．

1. **商品説明コーパスの用意**

   あるジャンル (「本」「パソコン」「化粧品」など) に属する複数の商品があり，その商品について説明した文書が存在すると仮定する．ここでの商品説明書は通販サイトにおいて商品を説明したテキストを想定する．以下，商品説明集合 $\mathcal{D} = \{d_1, \ldots, d_N\}$ と記述する．

2. **形態素解析と重要語候補の抽出**

   個々の商品説明 d_i を形態素解析し，重要語の候補となる単語を抽出する．形態素解析ツールとしては Janome を用いた．次に，助詞・助動詞・接続詞・記号などの機能語を除去し，名詞・動詞・形容詞などの内容語を抽出する．

3. **TF-IDF の算出**

   説明集合 $\mathcal{D}$ に対し，それに出現する単語 t の TF-IDF を算出する．TF-IDF は単語の重要度を定量化したスコアとみなせる．

4. **重要語の抽出**

   商品説明 d から TF-IDF スコアの上位 K 語を抽出し，重要語集合 T を作成する．本研究では $K = 3$ とし，重要語集合を $T = \{t_1, t_2, t_3\}$ と記す．

5. **レビューとの照合**

   レビュー r に対して同様に形態素解析を行い，T のいずれかがレビュー中に出現していれば $m(r, d) = 1$（言及あり），1 つも出現しなければ $m(r, d) = 0$（言及なし）とする．

本判定は「商品説明で強調されている特徴語がレビューに現れているか」に基づくため，配送・購入体験など商品内容以外を述べるレビューを「言及なし」と判定しやすい．

評価の有無の判定

評価の有無は，評価語辞書に含まれる語がレビュー中に出現するかにより判定する．本研究では評価語辞書として日本語評価極性辞書（用言編）[25] を用いる．判定の手続きを以下に示す．

1. レビュー r を形態素解析し，基本形（原形）へ正規化する．

2. 評価語辞書の単語集合 S と照合し，S に含まれる語が 1 語でも出現すれば $e(r) = 1$（評価あり），出現しなければ $e(r) = 0$（評価なし）とする．

この判定は単純である一方,「良い」「悪い」などの典型表現を取りこぼしにくい利点がある．

具体例

本ルールベース手法による疑似ラベル付与の動作例を示す．例として商品を「膝サポーター」とし，商品説明文集合から抽出された重要語を $T = \{$膝, サポーター, スポーツ$\}$ とする．

以下は「言及あり」と判定されるレビューの例である．

- 「最近，旦那さんはスポーツを始めたので，膝の負担を軽減するために購入しました」
  ⇒ 重要語（膝，スポーツ）が出現するため $m(r, d) = 1$（言及あり）
- 「サポーターは通気性が良く，ひざにフィットして動きやすいです」
  ⇒ 重要語（サポーター）が出現するため $m(r, d) = 1$（言及あり）

以下は「言及なし」と判定されるレビューの例である．

- 「発送が遅かったです」
  ⇒ 重要語が出現しないため $m(r, d) = 0$（言及なし）
- 「セールの日に購入しました」
  ⇒ 重要語が出現しないため $m(r, d) = 0$（言及なし）

次に，評価の有無の判定の例を示す．既に述べたように，本手法では単語を原形に変換した上で評価語辞書中の単語と照合しているため，以下の例における活用形の単語も原形が評価語辞書に含まれていれば評価語として検出される．

以下は「評価あり」と判定されるレビューの例である

- 「サイズ感もちょうどよく，とても良い商品だと思います」
  ⇒ 評価語（良い）が出現するため，$e(r) = 1$（評価あり）
- 「思ったより小さくて，持ち運びしやすい点が気に入っています」
  ⇒ 評価語（小さい）が出現するため，$e(r) = 1$（評価あり）

以下は「評価なし」と判定されるレビューの例である．

- 「膝に装着してウォーキングをしました」
  ⇒ 評価語が出現しないため，$e(r) = 0$（評価なし）

- 「購入後，数回使用しています」
  ⇒ 評価語が出現しないため，$e(r) = 0$（評価なし）

最終的には，$m(r, d)$ と $e(r)$ の組み合わせにより,「言及なし」「言及あり・評価なし」「言及あり・評価あり」の 3 値ラベルを決定する．

本手法の限界

本手法は実装が容易でラベル付けに要する計算コストが低い一方，(1) 同義語・言い換え・表記揺れ（例：「ひざ」 vs 「膝」），(2) 重要語が含まれないが商品内容を述べているレビュー，(3) 皮肉表現や否定表現など，評価語のマッチングでは検出できない評価表現を十分に捉えられない可能性がある．このようなルールベース手法の問題点が後述する LLM による疑似ラベル付与を検討する動機となった．

3.3.2 LLM を用いた疑似ラベル付与

ルールベース手法では，表記揺れや言い換え，文脈に依存する評価表現を十分に扱えないという課題があった．そこで，LLM でレビューに対する疑似ラベル付与を行う．

本手法では，LLM に対してレビューと商品説明を入力し,「商品への言及の有無」と「評価の有無」を同時に推定させる．その結果を用いて，レビューに対して 3 値の擬似ラベルを付与する．LLM としては，文脈理解能力が高く，日本語性能にも優れる大規模言語モデルである GPT-4o[1] を使用した．

プロンプト設計　LLM による疑似ラベル付与では，入力情報の与え方によって推定精度が大きく変化する．そこで本研究では，以下の 4 種類の入力を用意し，擬似ラベル付与精度の比較を行った．

1. レビュー＋商品説明
2. レビュー＋商品説明＋クラスの例
3. レビュー＋商品説明＋クラスの定義
4. レビュー＋商品説明＋クラスの例＋クラスの定義

それぞれのプロンプトを図 3.2，図 3.3，図 3.4，図 3.5 に示す．いずれのプロンプトにおいても，クラス定義と具体例を明示することで，LLM がタスクの意図を正確に理解できるよう工夫した．これら 4 つの入力による擬似ラベル付与の結果の比較については 4.3.1 項で報告する．

[1] OpenAI, GPT-4o, https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

```
# 命令
以下のレビュー文が、商品説明文の内容に言及しているか、また商品についての
評価を含んでいるかを判定してください。
以下の定義に従って、0、1、2 のいずれかの数字のみを出力してください。

## ラベルの例
- 【0: 言及なし】
  - 商品そのものについて触れていない
    - 例:「発送が早かった」「指定日に届いた」「セールだった」
  - 次の商品への期待や感想になっている
    - 例:「次も買います」「期待しています」
  - 自分の状態のみを説明している
    - 例:「2 人の子どもがいます」「現在腰痛を持っていて」

- 【1: 言及あり・評価なし】
  - 商品について言及しているが、良し悪しの評価がない
    - 例:「○○を購入しました」「○○のために活用したい」
  - 商品の素材や状態についての推測
    - 例:「○○の素材だと思う」
  - 他人の様子や主観的な推測
    - 例:「喜んでもらえた」「ぐっすり眠れてそう」

- 【2: 言及あり・評価あり】
  - 商品を使用した結果や満足・不満など評価が明確にある
    - 例:「生地が薄い」「使いやすい」「壊れやすい」
  - 商品の状態や品質について断定的に述べている
    - 例:「しっかりした作り」「色が写真と違う」
  - 特定の状況での使用結果や不具合を述べている

# 出力形式
0
1
2
のいずれか（数字のみ）

# 商品説明文
{product_description}

# レビュー
{review}
```

図 3.2: 擬似ラベル付与のためのプロンプト (レビュー＋商品説明)

```
# 命令
以下のレビュー文が、商品説明文の内容に言及しているか、また商品についての
評価を含んでいるかを判定してください。
以下の例に従って、0、1、2 のいずれかの数字のみを出力してください。

## ラベルの例
- 【0: 言及なし】
    - 例:「発送が早かった」「指定日に届いた」「セールだった」
    - 例:「次も買います」「期待しています」
    - 例:「2 人の子どもがいます」「現在腰痛を持っていて」

- 【1: 言及あり・評価なし】
    - 例:「○○を購入しました」「○○のために活用したい」
    - 例:「○○の素材だと思う」
    - 例:「喜んでもらえた」「ぐっすり眠れてそう」

- 【2: 言及あり・評価あり】
    - 例:「生地が薄い」「使いやすい」「壊れやすい」
    - 例:「しっかりした作り」「色が写真と違う」

# 出力形式
0
1
2
のいずれか（数字のみ）

# 商品説明文
{product_description}

# レビュー
{review}
```

図 3.3: 擬似ラベル付与のためのプロンプト（レビュー＋商品説明＋クラスの例）

```
# 命令
以下のレビュー文が、商品説明文の内容に言及しているか、また商品についての
評価を含んでいるかを判定してください。
以下の定義に従って、0、1、2 のいずれかの数字のみを出力してください。

## ラベルの定義
- 【0: 言及なし】
  - 商品そのものについて触れていない
  - 次の商品への期待や感想になっている
  - 自分の状態のみを説明している

- 【1: 言及あり・評価なし】
  - 商品について言及しているが、良し悪しの評価がない
  - 商品の素材や状態についての推測
  - 他人の様子や主観的な推測

- 【2: 言及あり・評価あり】
  - 商品を使用した結果や満足・不満など評価が明確にある
  - 商品の状態や品質について断定的に述べている
  - 特定の状況での使用結果や不具合を述べている

# 出力形式
0
1
2
のいずれか（数字のみ）

# 商品説明文
{product_description}

# レビュー
{review}
```

図 3.4: 擬似ラベル付与のためのプロンプト（レビュー＋商品説明＋クラスの定義）

```
# 命令
以下のレビュー文が、商品説明文の内容に言及しているか、また商品についての
評価を含んでいるかを判定してください。
以下の例と定義に従って、0、1、2 のいずれかの数字のみを出力してください。

## ラベルの例と定義
- 【0: 言及なし】
  - 商品そのものについて触れていない
    - 例:「発送が早かった」「指定日に届いた」「セールだった」
  - 次の商品への期待や感想になっている
    - 例:「次も買います」「期待しています」
  - 自分の状態のみを説明している
    - 例:「2 人の子どもがいます」「現在腰痛を持っていて」

- 【1: 言及あり・評価なし】
  - 商品について言及しているが、良し悪しの評価がない
    - 例:「○○を購入しました」「○○のために活用したい」
  - 商品の素材や状態についての推測
    - 例:「○○の素材だと思う」
  - 他人の様子や主観的な推測
    - 例:「喜んでもらえた」「ぐっすり眠れてそう」

- 【2: 言及あり・評価あり】
  - 商品を使用した結果や満足・不満など評価が明確にある
    - 例:「生地が薄い」「使いやすい」「壊れやすい」
  - 商品の状態や品質について断定的に述べている
    - 例:「しっかりした作り」「色が写真と違う」
  - 特定の状況での使用結果や不具合を述べている

# 出力形式
0
1
2
のいずれか（数字のみ）

# 商品説明文
{product_description}

# レビュー
{review}
```

図 3.5: 擬似ラベル付与のためのプロンプト（レビュー＋商品説明＋クラスの例＋クラスの定義）

3.4　有用性判定モデルの学習

本節では，前節で構築した擬似ラベル付きレビュー集合を用いて，有用性判定モデルを学習する方法について述べる．本研究では，事前学習済み言語モデルをベースとした分類モデルを構築し，レビューが商品内容に言及しているか，および商品に対する評価を含んでいるかを判定する．分類モデルの概要を図 3.3 に示す．レビューと商品説明を入力として与え，レビューを前述した 3 つのクラスのいずれかに分類する．

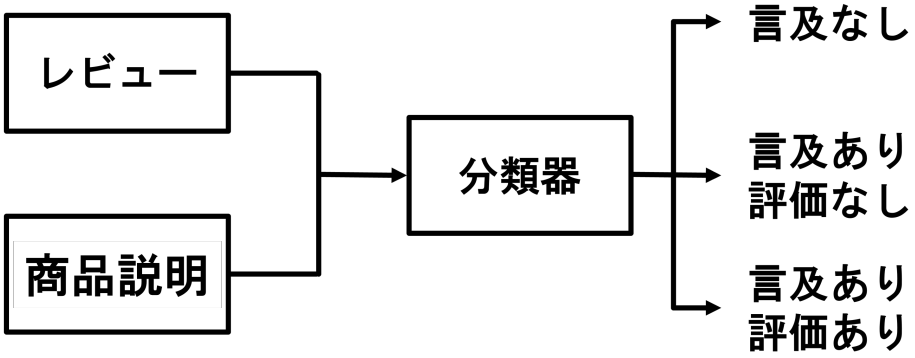


図 3.6: 有用性判定モデルの概要

3.4.1　事前学習済み言語モデルのファインチューニング

本研究では，有用性判定モデルとして，事前学習済み言語モデルである BERT と ModernBERT を用いる．

BERT は大規模なコーパスで事前学習されており，文脈を考慮した高い品質の文の抽象表現を獲得できることが知られている．本研究では東北大学が公開している日本語事前学習モデル `cl-tohoku/bert-base-japanese` を使用し，擬似ラベル付きレビュー集合を用いてファインチューニングを行う．

モデルへの入力は，レビューと商品説明を結合したテキストとし，以下の形式で与える．

`[CLS]` [レビュー] `[SEP]` [商品説明]

ここで，商品説明を同時に入力することで，レビュー文中に商品説明の内容が含まれているかどうかをモデルが文脈的に判断できるようにしている．これにより，単語一致に依存せず，言い換えや表現の違いを考慮した商品への言及判定が可能となる．

分類層としては，BERT の `[CLS]` トークンに対応する最終層の出力を用い，全結合層を介してクラス分類を行う．

ModernBERT は，BERT [29] を基盤としつつ，RoBERTa [30] や DeBERTa [31] などに代表される近年の改良手法の知見を取り入れた BERT 系モデルである．学習手法やモデル構造の最適化により，文脈表現能力の向上が図られており，比較的長い入力系列に対しても安定した表現獲得が可能である点が特徴である．

また，本研究で扱うレビューは，商品説明を付加することで入力系列が長くなる場合があるため，長文入力への対応能力はモデル選択において重要な要件となる．ModernBERT は，長文を含むテキスト分類タスクにおいて有効であることが報告されており [32]，本研究の設定とも親和性が高いと考えられる．

3.4.2 一段階方式と二段階方式

本研究では，有用性判定モデルの構成として，一段階方式と二段階方式の 2 種類の方式を検討した．それぞれの方式の概要を以下に示す．

一段階方式

図 3.7a に示すように，レビューと商品説明の組に対して，言及の有無と評価の有無を同時に判定し，以下の 3 クラスのいずれかに直接分類する．

- 言及なし
- 言及あり・評価なし
- 言及あり・評価あり

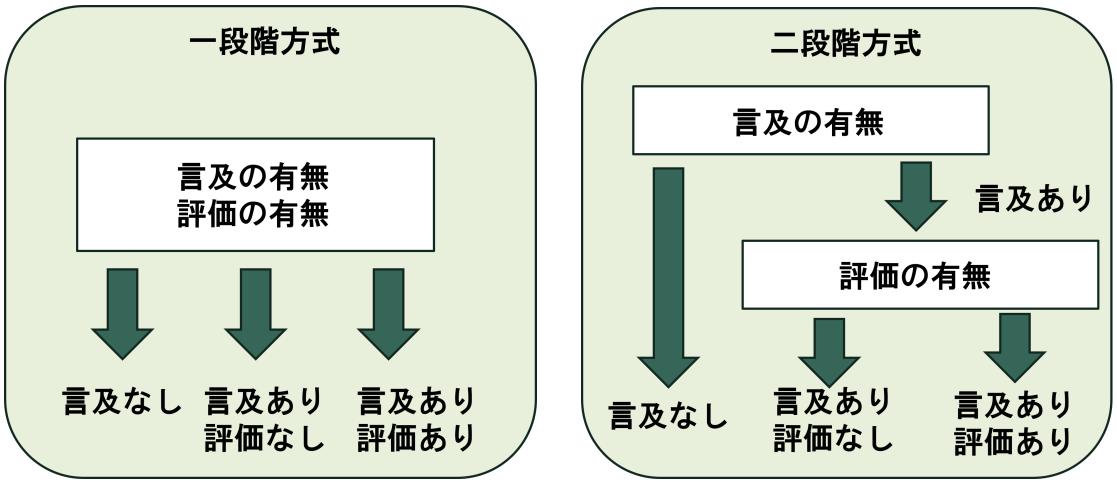
この方式では，単一の分類器によって 3 値分類を行うため，モデル構成が単純であり，言及あり・なしと評価あり・なしの判断を総合的に扱うことができる．

二段階方式

まずレビューが商品内容に言及しているかどうかを判定し，言及があると判定されたレビューのみを対象として，次に評価の有無を判定する．

この処理の流れを図 3.7b に示す．まず「言及の有無」を判定する分類器が存在し，言及ありと判定された場合のみ,「評価の有無」を判定する分類器へ入力される構成となる．

この方式では,「言及なし」のレビューを第 1 段階で除外するため，評価判定におけるノイズを低減できる利点がある．一方で，2 つの分類器を順に適用する必要があるため，誤りが後段に伝播する可能性がある．



(a) 一段階方式　　(b) 二段階方式

図 3.7: 一段階方式と二段階方式の比較

第4章　評価

本章では提案手法の評価実験について述べる．

4.1　データセット

本節では，使用したデータセットの概要と，訓練データおよびテストデータの作成方法について述べる．本研究では，レビューが商品に言及しているかどうか，および商品に対する評価を含んでいるかどうかを判定するため，大規模な商品レビューコーパスを用いて訓練データおよびテストデータを構築する．

4.1.1　訓練データ

訓練データの構築には，高度言語情報融合フォーラム ALAGIN を通じて提供されている楽天データセット [26] を利用した．本データセットは，楽天市場における商品情報およびユーザレビューを含むものであり，商品データは約 500GB，レビューデータは約 100GB の容量を有する．

商品データおよびレビューデータの構成

楽天市場データセットに含まれる商品データおよびレビューデータの情報をそれぞれ表 4.1，表 4.2 に示す．

表 4.1: 楽天データセットにおける商品データの構成

カラム名	内容
商品名	商品の名称
店舗コード	商品を販売する店舗の識別子
商品コード	店舗内での商品識別子
商品ページ URL	商品ページの URL
商品価格	商品の販売価格
商品ジャンル ID	楽天ジャンル ID
商品画像 URL	商品画像への URL
販売方法別説明文	販売方法ごとの説明文
商品説明	商品の詳細説明文

表 4.2: 楽天データセットにおけるレビューデータの構成

カラム名	内容
投稿者 ID	レビュー投稿者の識別子
店舗名	店舗の名称
店舗 ID	店舗の識別子
商品名	レビュー対象商品の名称
商品 ID	商品の識別子
商品ページ URL	商品ページの URL
商品ジャンル ID	楽天ジャンル ID
商品ジャンル ID パス	上位ジャンルを含むジャンル階層情報
使い道	商品の使用用途
目的	購入目的
頻度	使用頻度
評価ポイント	星評価（数値）
レビュータイトル	レビューのタイトル
レビュー内容	レビュー本文
参考になった数	有用性投票数
レビュー登録日時	レビュー投稿日時

使用カラムの抽出

本実験では楽天データセットのうちレビューと商品説明を使用した．すなわち，楽天データセットにおける全カラムのうち，必要最小限の情報のみを抽出して使用した．具体的に使用したカラムを表 4.3 に示す．

表 4.3: 本研究で使用したカラム

データ種別	使用カラム	用途
商品データ	商品名	レビューとの対応付け
	商品ページ URL	商品・レビューの紐付けキー
	商品ジャンル ID	ジャンル制御
	商品説明	言及判定の基準情報
レビューデータ	商品名	商品識別補助
	商品ページ URL	商品データとの紐付け
	商品ジャンル ID	ジャンル一致確認
	レビュータイトル	補助的文脈情報
	レビュー内容	分類対象テキスト

商品データとレビューデータの整形

楽天市場データセットでは，商品データとレビューデータが独立した形式で提供されているため，そのままでは商品とレビューの対応関係が明示されていない．そこで本研究では,「商品ページ URL」をキーとして，商品データとレビューデータを関連付け,「1 商品に対して複数のレビューが紐づく」形式へとデータを整形した．具体的には，商品データを基準とし，各商品に対してレビューを付与する構造を構築した．

使用データの条件設定

訓練データとして使用する商品およびレビューについては，データ品質の確保およびジャンル間の偏りを抑えることを目的として，いくつかの条件を設けた．

まず，十分なレビュー量が確保されている商品を対象とするため，1 商品につき 10 件以上のレビューが存在する商品に限定した．この条件を満たす商品の総数は 546,079 件である．各商品については，レビュー数の偏りを抑えるため，10 件以上存在するレビューの中からランダムに 5 件のみを抽出した．

また，レビューの長さによる影響を抑えるため，レビュー本文の文字数が 50 文字以上 150 文字以下のもののみを対象とした．参考のため，レビューの文字数の分布を図 4.1 に示す．これにより，極端に短いレビューや，冗長な説明を含む長文レビューを除外した．

上記に加え，本研究では，ジャンルごとの傾向差を考慮しつつ，十分なデータ数が確保可能なジャンルの商品に実験対象を限定した．楽天市場データセット全体におけるジャンルごとの商品数の分布を図 4.2 に示す．本データセットでは，ジャンルによって商品数に大きな偏りが存在することが分かる．そのため本研究では，

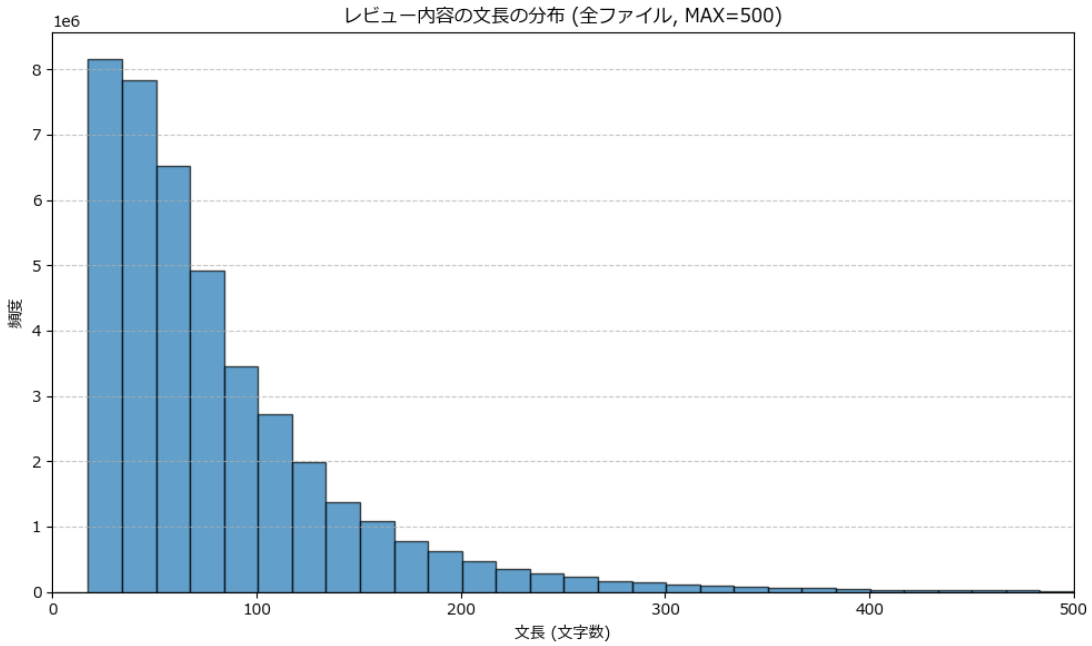


図 4.1: レビューの文字数の分布

十分なデータ数が確保可能であり，かつ多様なジャンルの商品を分析の対象とするため，実験に使用する商品ジャンルを以下の 10 個に限定した．

- 日用品
- スポーツ・アウトドア
- インテリア
- バッグ・小物
- 美容・コスメ
- 車用品・バイク用品
- 花・ガーデン
- レディースファッション
- 本
- 食品

これらの条件を満たす商品およびレビューを用いて，疑似ラベル付与を行い，訓練データを構築した．

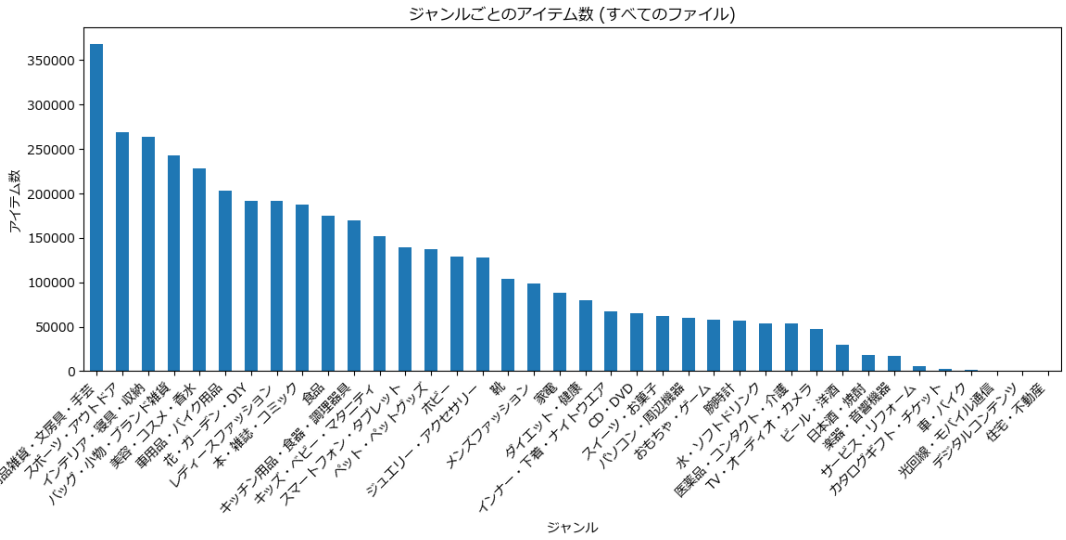


図 4.2: 楽天市場データセットにおけるジャンルごとの商品数分布

レビューの文分割と文単位でのラベル付与

楽天市場データセットに含まれるレビューは，1 件のレビューが複数の文から構成されている場合が多い．しかし，レビュー全体を 1 単位として扱うと，商品の内容に言及した文と言及していない文 (配送や購入体験に関する文) が混在し，有用性の判定が曖昧になる可能性がある．

そこで本研究では，レビューを句点「.」を区切り文字として文に分割し，1 文を最小単位として擬似ラベル付与を行った．具体的には，各レビューを文分割した後，それぞれの文 r に対して，対応する商品説明 d との組を作成し，言及の有無 $m(r, d)$ および評価の有無 $e(r)$ を判定した．

このように文単位でラベルを付与することで，同一レビュー内に含まれる「商品内容に関する文」と「商品内容に直接関係しない文」を区別でき，より正確な有用性判定モデルの学習が可能になる．

ルールベースの手法による疑似ラベル付与の訓練データ

ルールによる疑似ラベル付与では，大量のデータを用いて安定した学習を行うことを目的とし，比較的大規模な訓練データを構築した．先に述べた条件を満たす商品およびレビューの中から，10 ジャンルそれぞれについて均等にデータを抽出した．

ルールによる疑似ラベル付与に用いた訓練データの統計を表 4.4 に示す．

表 4.4: ルールによる疑似ラベル付与に用いた訓練データの統計

項目	データ数
レビュー数（全体）	100,000
商品数（全体）	20,000
ジャンル数	10
1 ジャンルあたりのレビュー数	10,000
1 ジャンルあたりの商品数	2,000

LLM を用いた疑似ラベル付与の訓練データ

LLM を用いた疑似ラベル付与では，ルールベース手法と比べてデータ規模は小さい．これは，ルールによる疑似ラベル付与と比較して推論にかかるコストが大きいためである．

LLM を用いた疑似ラベル付与に用いた訓練データの統計を表 4.5 に示す．

表 4.5: LLM を用いた疑似ラベル付与に用いた訓練データの統計

項目	データ数
レビュー数（全体）	20,000
商品数（全体）	4,000
ジャンル数	10
1 ジャンルあたりのレビュー数	2,000
1 ジャンルあたりの商品数	400

4.1.2 開発データ

開発データは，レビューに対して正解ラベルを人手でアノテーションすることで作成した．まず，前項で選定した 10 ジャンルそれぞれから，商品を 1 つずつ選択した．各商品について，5 件のレビューデータを抽出し，訓練データと同様に，レビューを句点「.」ごとに分割した．

作成したレビューに対して，1 名のアノテータがラベル付けを行った．各文に対して付与するラベルは,「言及なし」,「言及あり・評価なし」,「言及あり・評価あり」のいずれかである．

最終的に 150 件のレビューからなる開発データを作成した．表 4.6 に，作成した開発データのラベルの分布を示す.「言及あり・評価あり」が最も多く，一方で「言及なし」は比較的少数である．

以上の手続きで構築した開発データは，提案手法を設計する段階での予備実験で使用した．

表 4.6: 開発データにおけるラベル分布

ラベル	サンプル数
言及なし	36
言及あり・評価なし	44
言及あり・評価あり	70
合計	150

4.1.3 テストデータ

テストデータは，おおむね開発データと同じように，レビューに対して正解ラベルを人手でアノテーションすることで作成した．まず，4.1.1 項で選定した 10 ジャンルそれぞれから，商品を 1 つずつ選択した．各商品について，5 件のレビューデータを抽出し，訓練データと同様に，レビューを句点「.」ごとに分割した．

作成したレビューに対して，3 名のアノテータが独立にラベル付けを行った．各文に対して付与するラベルは,「言及なし」,「言及あり・評価なし」,「言及あり・評価あり」のいずれかである．アノテーション後，3 名のラベルの多数決を取り，最終ラベルを決定した．ただし，3 名全員が異なるラベルを付与した文はテストデータから除外した．

合計 162 件のレビューに対してアノテーションを行い，3 名のラベルがすべて異なるデータを除外したところ，最終的に 155 件のレビューからなるテストデータを作成した．このテストデータを用いて提案手法および比較手法の性能評価を行った．

表 4.7 に，作成したテストデータにおけるラベル分布を示す.「言及あり・評価あり」が最も多く，一方で「言及なし」は比較的少数である．

表 4.7: テストデータにおけるラベル分布

ラベル	サンプル数
言及なし	23
言及あり・評価なし	32
言及あり・評価あり	100
合計	155

4.2 実験条件

本節では，提案手法の有効性を検証するために実施した実験の条件について述べる．レビューに含まれる商品内容への言及の有無および評価の有無を判定する

【送料無料/あす楽対応】即日発送可能なカラフルなバラのフラワーギフト/誕生日/結婚記念日/発表会/送別会/退職祝い/生花 花束/バラの花バラ花束 送料無料 バラ 32 本 誕生日にバラをプレゼント【誕生日 発表会 記念日 お祝い 出産祝い 新築祝い 送別会 お見舞い】あす楽対応 即日発送 愛する方へ 生花 薔薇 バラの花束を フラワーギフト 母 姉 妹 バラの花

図 4.3: 長い商品名の例

有用性判定タスクを対象とし，疑似ラベル付与手法，事前学習済み言語モデル，有用性判定モデルの構成，および入力として与える情報の違いが性能に与える影響を検証した．

疑似ラベル付与手法としては，ルールベース手法と LLM を用いた手法の 2 種類を比較した．

事前学習済み言語モデルとしては BERT と ModernBERT を用いた．BERT として東北大学で公開している日本語 BERT モデル [12] を使用した．ModernBERT として，Hugging Face Transformers ライブラリ上で公開されている Japanese ModernBERT [13] を使用した．これらの事前学習済み言語モデルを擬似ラベル付きレビュー集合を用いてファインチューニングした．さらに，異なる乱数シードを用いてモデルのファインチューニングを複数回実行し，評価指標の平均を求めた．

有用性判定モデルの構成として，3 クラス分類を直接行う一段階方式と，「言及の有無」および「評価の有無」を段階的に判定する二段階方式を比較した．

さらに，入力情報の違いが分類性能に与える影響を分析するため，レビューのみを入力とするベースラインに加え，商品名および商品説明を組み合わせた複数の入力構成を検討した．加えて，入力系列の最大長を 256，512，1024 トークンのいずれかに設定し，有用性判定モデルの性能への影響を評価した．

入力形式　既に説明したように，提案手法ではレビューと商品説明を入力として与える．BERT や ModernBERT では，2 つの文を入力するときは

```
[CLS] sentence1 [SEP] sentence2 [SEP]
```

といった形式で 2 つの文を [SEP] トークンで分けて入力する．以下，最初の文を第 1 セグメント，2 番目の文を第 2 セグメントと呼ぶ．基本的には，レビューを第 1 セグメント，商品説明を第 2 セグメントに与える．

一方，楽天データセットでは商品説明の他に商品名の情報も収録されている．楽天データセットにおける商品やレビューは通販サイト「楽天市場」に掲載されたものであるが，楽天市場では多くのショップで長い商品名をつけている．したがって，商品名もまた商品に関する情報源となる．図 4.3 に楽天データセットにおける長い商品名の例を示す．

上記を考慮し，本研究では以下の 4 つの入力を与えたときの有用性判定モデルの性能を比較する．

- **レビュー**：レビューのみを入力する（1 セグメント）
- **レビュー＋商品名**：第 1 セグメントにレビューを，第 2 セグメントに商品名を入力する
- **レビュー＋商品説明**：第 1 セグメントにレビューを，第 2 セグメントに商品説明を入力する
- **レビュー＋商品名＋商品説明**：第 1 セグメントに「レビュー [SEP] 商品名」を，第 2 セグメントに商品説明を入力する

入力長がモデルの最大長を超えるときは，入力を先頭から最大長までのトークン列に切り詰めて入力する．レビューのみの条件では truncation=True により入力全体を上限長で切り詰めた．一方，2 セグメント入力では，レビュー（第 1 セグメント）の情報欠落を防ぐため，truncation=only_second を用い，第 2 セグメントのテキストを優先的に切り詰める設定とした．

4.3 実験結果

4.3.1 擬似ラベル付与手法の比較

本項では，ルールベースの手法と LLM を用いた手法によって与えられた擬似ラベルの品質を評価し，これらの 2 つの手法を比較する．

まず，ルールによる擬似ラベル付与手法を評価する．評価には，開発データを用いる．開発データの各文に対してルールベースの手法により「言及なし」「言及あり・評価なし」「言及あり・評価あり」の 3 値ラベルを付与し，人手ラベルとの (正解率) を算出する．

その結果，ルールによる疑似ラベル付与の正解率は 0.23 であった．この値は 3 値分類問題としては低い水準であり，ルールベース手法単体で高精度なラベル付与を行うことが困難であることを示している．ルールによる疑似ラベル付与の結果として得られたラベル分布を表 4.8 に示す．

表 4.8: ルールベース手法による疑似ラベルの分布

ラベル	サンプル数
言及なし	134
言及あり・評価なし	12
言及あり・評価あり	4
合計	150

表 4.9: LLM による擬似ラベル付与の評価 (プロンプト構成の比較)

プロンプト構成	正解率
レビュー＋商品説明	0.72
レビュー＋商品説明＋クラスの例	0.70
レビュー＋商品説明＋クラスの定義	0.69
レビュー＋商品説明＋クラスの例＋クラスの定義	0.74

「言及なし」に分類されたサンプルが 134 件と大半を占めており,「言及あり・評価なし」は 12 件,「言及あり・評価あり」は 4 件にとどまっていることが分かる．この結果は，ルールベース手法が多くのレビューを「言及なし」と判定する傾向を持つことを示している．特に，商品内容について言及していても，重要語集合に含まれない単語や言い換え表現を用いた場合には「言及なし」と誤判定される可能性が高い，また，評価表現についても，評価語辞書に含まれない主観的・文脈依存的な表現は検出できない．このような制約により,「言及あり・評価あり」クラスのサンプル数が極端に少なくなったと考えられる．

次に，LLM による疑似ラベル付与の評価を行う．表 4.9 に，4 つのプロンプトを用いたときの開発データにおける正解率を示す．結果を見ると,「レビュー＋商品説明」のみのベースライン（0.72）に対し,「クラスの例」のみの追加（0.70）や「クラスの定義」のみの追加（0.69）では，かえって正解率が低下した．これは，クラスの定義のみでは判断基準の抽象度が高く，一方でクラスの例のみでは情報が不十分であり，LLM による推論に悪影響を与えたためと考えられる．

これに対し,「クラスの例」と「クラスの定義」の双方を同時に与えた条件では，正解率が 0.74 となり，最も高い性能を示した．これは，定義によってタスクの明確な基準が示され，かつ例によってその具体的な適用事例が補完されたことで，定義と例の相乗効果が生まれ，LLM がタスクをより正確に理解できたためと解釈できる．

以上の結果から，LLM への与えるプロンプトとして「クラスの例」と「クラスの定義」を組み合わせる構成が有用性判定において最も有効であることが確認された．次に，このプロンプト構成（例＋定義）を基本設定として採用し，入力情報として「商品名」を加えることによる影響を検証する．具体的には，以下の 2 つの入力構成の比較実験を行う．

1. レビュー＋商品説明＋クラスの例＋クラスの定義

2. レビュー＋商品名＋商品説明＋クラスの例＋クラスの定義

1 番目の入力に対するプロンプトは図 3.5 に示した通りである．2 番目の入力に対するプロンプトを図 4.4 に，示す．

表 4.10 に，入力構成の違いによる有用性判定精度を示す．

```
# 命令
以下のレビュー文が、商品名と商品説明文の内容に言及しているか、また商品についての評価を含んでいるかを判定してください。
以下の例と定義に従って、0、1、2 のいずれかの数字のみを出力してください。

## ラベルの例と定義
- 【0: 言及なし】
  - 商品そのものについて触れていない
    - 例:「発送が早かった」「指定日に届いた」「セールだった」
  - 次の商品への期待や感想になっている
    - 例:「次も買います」「期待しています」
  - 自分の状態のみを説明している
    - 例:「2 人の子どもがいます」「現在腰痛を持っていて」

- 【1: 言及あり・評価なし】
  - 商品について言及しているが、良し悪しの評価がない
    - 例:「○○を購入しました」「○○のために活用したい」
  - 商品の素材や状態についての推測
    - 例:「○○の素材だと思う」
  - 他人の様子や主観的な推測
    - 例:「喜んでもらえた」「ぐっすり眠れてそう」

- 【2: 言及あり・評価あり】
  - 商品を使用した結果や満足・不満など評価が明確にある
    - 例:「生地が薄い」「使いやすい」「壊れやすい」
  - 商品の状態や品質について断定的に述べている
    - 例:「しっかりした作り」「色が写真と違う」
  - 特定の状況での使用結果や不具合を述べている

# 出力形式
0
1
2
のいずれか（数字のみ）

# 商品名
{product_name}

# 商品説明文
{product_description}

# レビュー
{review}
```

図 4.4: 擬似ラベル付与のためのプロンプト（レビュー＋商品説明＋クラスの例＋クラスの定義）

表 4.10: LLM による擬似ラベル付与の評価 (入力構成の比較)

入力構成	Accuracy
レビュー＋商品説明＋クラスの例＋クラスの定義	0.74
レビュー＋商品名＋商品説明＋クラスの例＋クラスの定義	0.76

2 つの入力構成を使用したときの擬似ラベル付与の正解率を表 4.10 に示す．先の実験で最も有効であった「レビュー＋商品説明＋クラスの例＋クラスの定義」の構成（正解率 0.74）に対し，さらに「商品名」を追加した条件では，正解率が 0.76 へと向上した．この結果は，商品名が有用性判定において有効な情報源として機能したことを示している．先にも述べたように，本実験で使用した楽天データセットにおいて，商品名は単なる名称にとどまらず，商品の主要な特徴やセールスポイントなどのキーワードが羅列されている場合が多い．そのため，商品説明文だけでは捉えきれなかった，あるいは長文の説明の中に埋もれてしまっていた重要な特徴語が，商品名によって端的に提示されたことで，LLM がレビューと商品の関連性をより正確に捉えられるようになったと考えられる．

以上の結果から疑似ラベルを作成するプロンプトには最も精度の良かった「レビュー＋商品名＋商品説明＋クラスの例＋クラスの定義」を用いる．

4.3.2 2 つの擬似ラベル付与手法で作成されたデータセットによって学習されたモデルの比較

本実験では，分類モデルとして BERT を用いて，ルールベース手法（M_{rule}）と，LLM を用いた手法（M_{LLM}）により生成した擬似ラベルを用いて学習した分類モデルの性能を比較した．評価指標として，各クラスに対する Precision，Recall，F1 スコア，および全体の正解率（Accuracy）を用いた．表 4.11 および表 4.12 に擬似ラベル付与手法ごとの分類結果を示す．

表 4.11: ルールベース手法で作成した擬似ラベルデータセットから学習したモデルの評価

M_{rule}			
分類クラス	Precision	Recall	F1
言及なし	0.24	0.78	0.37
言及あり・評価なし	0.57	0.23	0.33
言及あり・評価あり	0.57	0.09	0.16
正解率		0.32	

表 4.12: LLM を用いた手法で作成した擬似ラベルデータセットから学習したモデルの評価

	M_{LLM}		
分類クラス	Precision	Recall	F1
言及なし	0.53	0.71	0.61
言及あり・評価なし	0.66	0.62	0.64
言及あり・評価あり	0.86	0.78	0.82
正解率		0.71	

ルールベース手法では,「言及あり・評価あり」クラスにおける Recall が低く,正解率も 0.32 と低い値にとどまった．一方，LLM を用いた手法では，すべてのクラスにおいて Precision および Recall が向上し，正解率は 0.71 に達した．この結果から，文脈理解を考慮できる LLM による疑似ラベル付与が，より高品質な擬似ラベルを生成できていることが確認された．

以上の結果から，LLM を用いた疑似ラベル付与によって得られた訓練データを用いる方が高い分類性能を示すことを確認した．以降の実験では，LLM による擬似ラベル付きデータセットを訓練データとして用いる．

4.3.3 一段階方式と二段階方式の比較

本節では，有用性判定モデルの構成として，一段階方式と二段階方式を比較し，分類性能に与える影響を検証する．一段階方式は，レビューから直接 3 クラス分類を行う単一モデルである．一方，二段階方式は，3 クラス判定を (1) 言及の有無の判定，(2) 評価の有無の判定という 2 つの二値分類問題に分割して逐次的に推定する方式である．

両方式の比較にあたっては，擬似ラベル付与手法の比較（§4.3.1）の結果を踏まえ，LLM により付与した擬似ラベルを学習データとして用いた．また，方式の違いのみを評価するため，入力はレビューのみとし，事前学習済み言語モデルは BERT を共通に用いた．最長トークン長（128），最適化関数（AdamW），学習率（2×10^{-5}），エポック数（3），およびクラス不均衡への対応としてのアンダーサンプリング等の学習条件は，両方式で同一とした．評価指標として，各クラスの Precision，Recall，F1 スコア，および全体の正解率を用いた．

表 4.13 および表 4.14 に一段階方式および二段階方式による分類結果を示す．

表 4.13: 一段階方式による有用性判定結果

分類クラス	Precision	Recall	F1
言及なし	0.53	0.71	0.61
言及あり・評価なし	0.66	0.62	0.64
言及あり・評価あり	0.86	0.78	0.82
正解率		0.71	

表 4.14: 二段階方式による有用性判定結果

分類クラス	Precision	Recall	F1
言及なし	0.41	0.86	0.55
言及あり・評価なし	0.66	0.38	0.48
言及あり・評価あり	0.84	0.72	0.78
正解率		0.63	

表 4.13 より，一段階方式の正解率は 0.71 であり，二段階方式の正解率である 0.63 ををを上回った．クラス別にみると，一段階方式は「言及あり・評価なし」および「言及あり・評価あり」においてF1 がそれぞれ 0.64，0.82 と高く，言及の有無と評価表現の有無を同一モデル内で整合的に捉えられていることが分かる．一方，二段階方式では「言及なし」の Recall が 0.86 と高いものの,「言及あり・評価なし」の Recall が 0.38 と低く，このクラスの F1 も 0.48 にとどまった．

この性能差は，二段階方式に内在する誤りの伝播の影響によって説明できる．二段階方式では，第 1 段階で「言及あり」であるサンプルを誤って「言及なし」と判定した場合，そのサンプルは第 2 段階に入力されないため，最終的に必ず「言及なし」として出力される．すなわち，第 1 段階の誤りは後段で訂正不可能であり,「言及あり・評価なし」「言及あり・評価あり」の各クラスの Recall を低下させる．特に，本実験では「言及あり・評価なし」の Recall が大きく低下しており，このクラスに属するサンプルの一部が第 1 段階で「言及なし」クラスに誤分類されていることが原因のひとつと考えられる．

さらに，二段階方式では第 2 段階の学習・推論対象が『「言及あり・評価なし」と「言及あり・評価あり」にサンプルが限定された部分集合』となるため，相対的に訓練データ量が少なくなり，一段階方式のように 3 クラスを同時に識別する学習に比べて，分類クラスの決定境界が不安定になりやすい点も不利に働くと考えられる．一方で，一段階方式は 3 クラスを直接最適化するため，言及判定と評価判定を同時に学習し，両者の相関をモデル内部に取り込める利点がある．

以上より，今回の実験においては，二段階方式よりも一段階方式の方が高い性能を示すことを確認した．以降の実験では，一段階方式を基本構成として採用する．

4.3.4 BERT と ModernBERT の比較

本節では，分類器として用いる事前学習済み言語モデルの違いがレビューの有用性判定性能に与える影響を検証する．既に述べたように，比較対象として，日本語 BERT [27] と，長文入力を想定して設計された ModernBERT[28] を用いる．

モデル以外の条件，すなわち入力データ，クラス不均衡への対応（アンダーサンプリング），最適化関数（AdamW），学習率（2×10^{-5}），早期終了の設定は同一とする．入力は，レビューのみ，レビュー＋商品説明，レビュー＋商品名，レビュー＋商品名＋商品説明の 4 つとし，Accuracy と Macro-F1 を算出した．

表 4.15: BERT と ModernBERT の比較

モデル	指標	レビューのみ	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
BERT	Accuracy	0.696	0.708	0.683	0.702
	Macro-F1	0.640	0.662	0.633	0.662
ModernBERT	Accuracy	0.743	0.750	0.737	0.717
	Macro-F1	0.699	0.704	0.681	0.655

実験結果を表 4.15 に示す．大部分の設定において ModernBERT は BERT を上回る傾向を示した．特に「レビューのみ」および「レビュー＋商品説明」を入力としたとき，Accuracy と Macro-F1 の双方で一貫した改善が確認された．これは，ModernBERT がレビュー文中の商品特徴の記述，使用感の描写，評価表現などをより適切に捉え，有用性判定の性能向上につながったと考えられる．

一方で，「レビュー＋商品名＋商品説明」のように入力情報量が多い設定では，ModernBERT の優位性は Accuracy では維持されるものの，Macro-F1 の改善幅は小さく，条件によっては BERT と同程度となった．この要因として，(1) 商品名および商品説明がレビューと直接関係しない情報を含み得ること，(2) 商品説明が定型的で冗長な表現を含み，判定に有効な特徴が相対的に薄まることが考えられる．すなわち，入力の追加は情報量を増やす一方で，ノイズや冗長性を導入し，クラス間の識別能力が必ずしも改善されないことを示している．

以上より，ModernBERT は BERT に比べて高い性能を示し，レビューのみ，もしくはレビューと商品説明または商品名を入力とする場合など，比較的入力文が短い設定において，有用性判定に有効な文脈情報をより適切に捉えられることが示唆された．この結果を踏まえ，以降の実験では，分類モデルとして ModernBERT を用いる．

4.3.5 提案手法の評価

前項までの比較実験により，(1) 擬似ラベル付与手法としては，ルールベースよりも LLM を用いた方法の方が安定して高い分類性能につながること，(2) 分類器の構成としては，二段階方式よりも一段階方式の方が前段の判定誤りが後段に影

響する誤り伝播を回避でき，性能および安定性の両面で有利であること，(3) 事前学習済み言語モデルとしては，BERT よりも ModernBERT の方が本タスクにおいて高い性能を示す傾向があること，を確認した．

これらの結果を踏まえ，本研究では，LLM により付与した擬似ラベル付き訓練データを用い，一段階方式による 3 クラス分類として ModernBERT をファインチューニングする構成を基本構成として採用する．本項では，この構成において，(1) 入力として与える情報の違い，(2) 入力系列長の設定，(3) 学習手順の違い，が性能および安定性に与える影響を詳細に分析する．

モデル学習設定

本実験では以下の 2 つの設定でモデルの学習を行う．

- **学習設定 (ES あり)**
  分類モデルを学習 (ファインチューニング) する際，最大エポック数を 3 と設定し，Early Stopping を適用する．

- **学習設定 (ES なし)**
  分類モデルを学習 (ファインチューニング) する際，エポック数を 5 に設定する．Early Stopping は適用しない．

実験当初は「設定 (ES あり)」の設定で実験を行った．しかし，モデルの学習と評価を 3 回試行したところ，評価指標のばらつきが大きかった．また，多くの場合に Early Stopping の結果エポック数が 1 で学習が完了し，かつ訓練データに対する損失が十分に下がらないことを確認した．そこで,「設定 (ES なし)」の設定で再実験を行った．この実験では，訓練データに対する損失を十分に下げることを目的とし，Early Stopping を使用しなかった．学習曲線を確認したところ，概ね 5 エポック目までに学習損失が収束し，モデルが十分に学習できていることを確認した．

結果と考察

学習設定 (ES あり) と学習設定 (ES なし) の 2 つのケースについて，Accuracy と Macro-F1 の平均および分散を報告し，考察を行う．

学習設定 (ES あり) の Accuracy の平均を表 4.16 に，Macro-F1 の平均を表 4.18 に示す．また，本実験ではモデルの学習と評価を 3 回行っているが，学習結果のばらつきを示すため，Accuracy の分散を表 4.17 に，Macro-F1 の分散を表 4.19 に示す．

表 4.16 より，最大トークン長 256 では「レビュー＋商品名＋商品説明」が Accuracy=0.732 と最も高く，次いで「レビュー＋商品名」（0.730）が高い．一方，最

表 4.16: 学習設定 (ES あり) での有用性判定の実験結果 – Accuracy の平均

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	0.714	0.722	0.730	0.732
512	0.722	0.689	0.714	0.718
1024	–	0.716	0.726	0.716

表 4.17: 学習設定 (ES あり) での有用性判定の実験結果 – Accuracy の分散

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	6.73×10^{-4}	6.24×10^{-4}	7.24×10^{-5}	2.00×10^{-4}
512	1.13×10^{-3}	6.20×10^{-4}	4.64×10^{-4}	9.28×10^{-4}
1024	–	5.13×10^{-4}	1.86×10^{-3}	4.24×10^{-4}

表 4.18: 学習設定 (ES あり) での有用性判定の実験結果 – F1-score の平均

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	0.667	0.670	0.677	0.667
512	0.679	0.643	0.657	0.670
1024	–	0.662	0.668	0.665

表 4.19: 学習設定 (ES あり) での有用性判定の実験結果 – F1-score の分散

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	4.33×10^{-4}	7.00×10^{-4}	3.33×10^{-5}	4.33×10^{-4}
512	1.20×10^{-3}	3.24×10^{-5}	1.03×10^{-3}	7.00×10^{-4}
1024	–	1.67×10^{-4}	1.12×10^{-3}	7.42×10^{-4}

大トークン長 512 では「レビューのみ」が 0.722 と最も高く,「レビュー＋商品説明」は 0.689 と低下した．最大トークン長 1024 では「レビュー＋商品名」が 0.726 と最も高いが，商品説明を含む構成（0.716〜0.716）は大きな改善には至っていない．これらの結果は，商品情報の付加が常に有効とは限らず，特に商品説明の付加は，条件によって性能低下も起こし得ることを示している．

表 4.18 に示す Macro-F1 でも Accuracy と同様の傾向が見られる．最大トークン長 256 では「レビュー＋商品名」が 0.677 と最も高く,「レビューのみ」および「レビュー＋商品名＋商品説明」は 0.667 に留まる．最大トークン長 512 では「レビューのみ」が 0.679 と最も高い一方,「レビュー＋商品説明」は 0.643 と大きく低下した．最大トークン長 1024 でも「レビュー＋商品名」が 0.668 と最大であるが，商品説明を含む構成（0.662〜0.665）との差は小さい．このことから，商品説明を追加して入力情報量を増やしても，クラス間のバランスを考慮した指標（Macro-F1）では改善が限定的であることが分かる．

複数回の学習における有用性判定モデルの性能のばらつきは，表 4.17 および表 4.19 から確認できる．例えば最大トークン長 256 において,「レビュー＋商品名」は Accuracy の分散が 6.73×10^{-4}，Macro-F1 の分散が 3.33×10^{-5} と小さく，比較的安定している．一方で最大トークン長 1024 の「レビュー＋商品名」は Accuracy の分散が 1.86×10^{-3} と大きく，長い入力条件では学習ごとの差が大きくなることが示唆される．Macro-F1 においても，最大トークン長 512 の「レビューのみ」は分散が 1.20×10^{-3} と大きく，同一条件でも学習の揺らぎが発生している．これらの分散の大きさは，Early Stopping が適用されるタイミングや初期パラメタの設定が評価値の変動要因となり得ることを示している．

上記の課題に対して，学習設定 (ES なし) ではエポック数を統一し，訓練データの損失を十分に下げた状況の下で，入力構成の影響をより公平に比較した．学習設定 (ES なし) における Accuracy の平均を表 4.20 に，Macro-F1 の平均を表 4.22 に示し，分散をそれぞれ表 4.21，表 4.23 に示す．

表 4.20 より，最大トークン長 256 では「レビュー＋商品名」が 0.742 と最も高く，学習設定 (ES あり) の結果と一致している．一方で，学習設定 (ES あり) で高かった「レビュー＋商品名＋商品説明」は 0.714 と低く，商品説明の付加が必ずしも有効でないことが，エポック数を揃えた条件でも再確認できる．最大トークン長 512 では「レビューのみ」が 0.730 と最も高く，学習設定 (ES あり) と同様に，商品説明の付加（0.711）は改善に寄与していない．最大トークン長 1024 でも「レビュー＋商品名」が 0.724 と最大であり，商品説明を含む構成（0.714，0.705）はそれを上回らない．

Macro-F1 についても，表 4.22 より，最大トークン長 256 では「レビュー＋商品名」が 0.695 と最も高い．最大トークン長 512 では「レビューのみ」が 0.680 と最も高く，最大「レビュー＋商品名」は 0.660 に留まる．最大トークン長 1024 では「レビュー＋商品名」が 0.675 と最大であり，商品説明の追加（0.665，0.660）は改善幅が小さい．以上より，学習設定 (ES なし) においても,「商品名の付加は

表 4.20: 学習設定 (ES なし) での有用性判定の実験結果 – Accuracy の平均

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	0.736	0.721	0.742	0.714
512	0.730	0.711	0.705	0.717
1024	–	0.714	0.724	0.705

表 4.21: 学習設定 (ES なし) での有用性判定の実験結果 – Accuracy の分散

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	1.92×10^{-5}	7.69×10^{-5}	4.84×10^{-4}	3.08×10^{-4}
512	1.92×10^{-5}	1.73×10^{-4}	1.73×10^{-4}	1.92×10^{-5}
1024	–	7.69×10^{-5}	1.92×10^{-5}	1.73×10^{-4}

表 4.22: 学習設定 (ES なし) での有用性判定の実験結果 – F1-score の平均

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	0.690	0.675	0.695	0.675
512	0.680	0.655	0.660	0.675
1024	–	0.665	0.675	0.660

表 4.23: 学習設定 (ES なし) での有用性判定の実験結果 – F1-score の分散

トークン長	レビュー	レビュー＋商品説明	レビュー＋商品名	レビュー＋商品名＋商品説明
256	2.00×10^{-4}	5.00×10^{-5}	4.50×10^{-4}	5.00×10^{-5}
512	0.000	4.50×10^{-4}	2.00×10^{-4}	5.00×10^{-5}
1024	–	5.00×10^{-5}	5.00×10^{-5}	2.00×10^{-4}

有効となる場合が多い一方，商品説明の付加は一貫した改善につながらない」という結論が得られた．

学習設定 (ES なし) の分散は表 4.21 および表 4.23 で確認できる．まず Accuracy では，最大トークン長 256 の「レビューのみ」が 1.92×10^{-5} と小さい一方,「レビュー＋商品名」は 4.84×10^{-4} と相対的に大きい．これは，商品名付加が平均性能を押し上げる一方で，有用性判定モデルの性能が不安定になることを意味する．最大トークン長 512 では,「レビュー＋商品名」および「レビュー＋商品説明」の分散がいずれも 1.73×10^{-4} であり，付加情報の種類によっても揺らぎが大きくなっている．最大トークン長 1024 では「レビュー＋商品名」が 1.92×10^{-5} と小さい一方,「レビュー＋商品名＋商品説明」は 1.73×10^{-4} と大きく，情報量の多い構成ほど学習が安定しない可能性が示唆される．

Macro-F1 でも，表 4.23 に示すように，最大トークン長 512 の「レビューのみ」は分散が 0.000 と小さい一方,「レビュー＋商品説明」は 4.50×10^{-4} と大きい．また，最大トークン長 256 の「レビュー＋商品名」は 2.00×10^{-4} と比較的大きく，平均性能が高い構成であっても，安定性の観点では劣ることがある．これらの結果は，提案手法の運用を考える際に，単純な平均性能だけでなく，学習の安定性（分散）も合わせて入力構成を選択する必要があることを示している．

学習設定 (ES あり)（表 4.16，表 4.18）および学習設定 (ES なし)（表 4.20，表 4.22）のいずれにおいても，256 → 512 → 1024 と系列長を増やしても性能が単調に向上する傾向は確認できなかった．この事実は,「長文を入力として与えられること」自体よりも，追加される情報がタスクに対して有効であるか，またモデルがそれを適切に利用できるかが重要であることを示している．特に商品説明は，定型文や冗長表現を含むことが多く，レビューが言及している商品の特徴と一致しない場合も多い．このような状況では，最大系列長を増やして商品説明を削除することなく入力できても，判定に有用な手がかりが増えるとは限らない．したがって，本タスクでは，最大系列長の拡大よりも，レビューと商品情報の対応関係を強める入力設計が今後の性能向上に重要であると考えられる．

以上より，提案手法（LLM 擬似ラベル＋一段階方式＋ ModernBERT）は，入力構成および最大系列長の設定により性能が変化し，さらに安定性にも差が生じることが，学習設定 (ES あり) の平均・分散（表 4.16〜表 4.19）と学習設定 (ES なし) の平均・分散（表 4.20〜表 4.23）の双方から確認された．特に，平均性能の観点では「レビュー＋商品名」が有利となる場面が多い一方，分散の観点では条件によって揺らぎが生じ得るため，運用目的に応じて性能と安定性を併せて判断する必要がある．

4.3.6 大規模言語モデルとの比較

大規模言語モデルによるゼロショットの有用性判定と，本研究で提案した擬似ラベルでファインチューニングした言語モデルとの比較を行った．具体的には，

ChatGPT（GPT-4o）に対して，レビューおよび商品情報を入力し，言及なし，言及あり・評価なし，言及あり・評価ありを「言及なし」,「言及あり・評価なし」,「言及あり・評価あり」のどれに該当するかを推定させた．

表 4.24: 大規模言語モデル（ChatGPT）と提案手法の比較

モデル	入力情報	正解率
ChatGPT (GPT-4o)	レビューのみ	0.776
	レビュー＋商品名	0.776
	レビュー＋商品説明	**0.796**
	レビュー＋商品名＋商品説明	0.776
提案手法 (Best)	レビュー＋商品名	0.742

実験結果を表 4.24 に示す．GPT-4o によるゼロショット手法は提案手法の正解率を上回った．特に,「レビュー＋商品説明」を与えた場合に最も高い正解率が得られており，LLM が商品説明を含む文脈を有効に利用できることが確認できる．

一方，提案手法の最高正解率は 0.742 であり，LLM の正解率には及ばなかった．しかし，両者の正解率の差はそれほど大きくない．また，ModernBERT は GPT-4o などの大規模言語モデルと比較してパラメタ数が大幅に少なく，計算コストの面で効率的かつ軽量なモデルであるという利点がある．

一方で，提案手法における ModernBERT を用いた分類モデルでは，入力構成によって性能に差は見られるものの，商品説明の付加が常に性能向上につながるわけではなかった．この差異は，LLM が事前学習段階で獲得した広範な知識や推論能力により，冗長な商品説明からも有用な情報を抽出できるのに対し，軽量な分類モデルでは，必ずしも同様の処理が行えないことに起因すると考えられる．

第5章　おわりに

5.1　まとめ

本研究では，EC サイトにおける商品レビューを対象として，レビューが商品内容に言及しているかどうか，および評価を含んでいるかどうかの観点から，レビューの有用性を 3 クラス（言及なし，言及あり・評価なし，言及あり・評価あり）に分類する手法を提案した．

従来のレビュー有用性判定研究では，レビューの長さや感情極性，あるいは他のユーザによる投票数などの外的指標に依存する手法が多く，レビューが商品そのものについてどの程度具体的な情報を提供しているかという観点は，必ずしも明示的に扱われていなかった．本研究は，商品への言及および評価の有無という観点からレビューの有用性を評価する点に特徴がある．

有用性判定モデルを学習するための訓練データを構築するにあたり，大規模な人手アノテーションが困難であるという実運用上の制約に対し，弱教師あり学習の枠組みを採用した．具体的には，(1) ルールベースおよび大規模言語モデル（LLM）を用いた擬似ラベル付与により訓練データを自動構築し，(2) 事前学習済み言語モデルをファインチューニングすることで，有用性判定モデルを学習した．入力として，レビューと商品説明を与え，レビューが商品内容に言及しているか，および評価を含んでいるかを判定できる有用性判定モデルを学習した．

評価実験では，まず擬似ラベル付与手法の比較を行い，LLM を用いた擬似ラベルが，ルールベース手法と比べてより高い分類性能を持つモデルを学習できることを確認した．次に，有用性判定モデルの構成として，一段階方式（3 クラスを直接分類）と二段階方式（言及判定後に評価判定）を比較し，二段階方式では後段への誤り伝播が生じやすいのに対し，一段階方式の方が全体として安定した性能を示すことを明らかにした．さらに，分類器として BERT と ModernBERT を比較した結果，ModernBERT がおおむね高い性能を示す傾向が確認された．

これらの結果を踏まえ，本研究では,「LLM による擬似ラベル付与」,「一段階方式による 3 クラス分類」,「ModernBERT の活用」を組み合わせた構成を提案手法として採用し，その性能評価を行った．その際，レビューへの商品情報（商品名・商品説明）の付加方法や，入力系列長（256，512，1024 トークン）の違いが性能に与える影響を詳細に分析した．また，モデルの学習する際の条件として，Early

Stopping を用いる設定と用いない設定とで実験を行った．

実験の結果，入力系列長を増加させても必ずしも性能が向上するとは限らず，レビューと商品情報の関係性や，入力情報の冗長性が性能に影響を与えることが示唆された．このことは，長文入力を許容できるモデルであっても，情報量の増加が常に有効に機能するわけではないことを示している．

ChatGPT（GPT-4o）を用いた直接推論の性能を確認したところ，本研究で学習したモデルよりも高い精度が得られた．一方で，ChatGPT は推論コストや再現性，運用環境の制約といった点で実サービスへの導入が容易とは限らない．これに対し，本研究で構築したモデルは，比較的軽量な構成で安定した推論が可能であり，ローカル環境やシステム組み込みを前提とした実運用上の利点を有している．

以上より，本研究は，商品言及と評価の観点からレビュー有用性を整理し，事前学習済み言語モデルを利用することでレビューの有用性判定を実現する現実的な手法を提案した．特に，人手ラベルに依存せずに一定水準の性能を達成できる点は，大規模データを扱う実サービスにおいて有用であると考えられる．

5.2 今後の課題

本研究には，さらなる性能向上と実運用を見据えた拡張に向けて，いくつかの課題が残されている．

第一に，擬似ラベルの品質向上である．LLM による擬似ラベルはルールベース手法より高品質である一方，判断の揺らぎや誤りが完全に排除されるわけではない．今後は，少量の人手ラベルを用いた検証や補正，あるいは擬似ラベルの信頼度を考慮した学習手法を導入することで，訓練データの品質をさらに高めることが必要である．

第二に，入力設計のさらなる最適化である．本研究では，商品名や商品説明の付加が有効に働く場合がある一方で，与える入力の冗長性が増したり，ノイズとして作用する可能性も確認された．今後は，商品説明の要約や重要文抽出，レビューと商品情報の関連部分を強調する仕組みなどを導入し，必要な情報を効率的に利用する設計を検討する必要がある．

第三に，評価データおよび評価指標の拡充である．本研究のテストデータは人手アノテーションにより構築したが，比較的小規模であり，また評価対象とした商品のジャンルも限られていた．今後は，より多様な商品カテゴリやレビュー特性を含むデータを用いた評価が必要である．これに加え，実際の購買行動やユーザ満足度との関係を考慮した評価指標の導入も検討したい．

最後に，実運用を想定した出力結果の扱いも重要な課題である．予測結果の信頼度提示や，誤分類時の影響を考慮した運用ルールを整備することで，レビュー有用性判定を実サービスでより安全かつ効果的に活用できると考えられる．

以上の課題に取り組むことで，商品言及に基づくレビュー有用性判定手法をより実用的な場面に活用することが可能になると考えられる．

関連図書

[1] Kuan, K. Y. K., Hui, K.-L., Prasarnphanich, P., and Lai, H.-Y. What makes a review voted? An empirical investigation of review voting in online review systems. *Journal of the Association for Information Systems*, vol. 16, pp. 48–71, 2015.

[2] McAuley, J. and Leskovec, J. Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems (RecSys)*, pp. 165–172, 2013.

[3] Singh, Jyoti P., Irani, Seda, Rana, Nripendra P., Dwivedi, Yogesh K., Saumya, S., and Roy, Pradip K. Predicting the "helpfulness" of online consumer reviews. Journal of Business Research, Vol. 70, pp. 346–355, 2017.

[4] 高島 侑里，青野 雅樹. 化粧品レビューサイトにおけるクチコミの有用性判定. 言語処理学会第 23 回年次大会発表論文集，pp. 799–802, 2017.

[5] Hong, Hong, Xu, Di, Wang, G. Alan, and Fan, Weiguo. Understanding the determinants of online review helpfulness: A meta-analytic investigation. Decision Support Systems, Vol. 102, pp. 1–11, 2017.

[6] Qin, Jindong, Zheng, Pan, and Wang, Xiaojun. Comprehensive helpfulness of online reviews: A dynamic strategy for ranking reviews by intrinsic and extrinsic helpfulness. Decision Support Systems, Vol. 163, 113859, 2022.

[7] Fan, Miao, Feng, Chao, Guo, Lin, Sun, Mingming, and Li, Ping. Product-aware helpfulness prediction of online reviews. In Proceedings of The World Wide Web Conference, pp. 2715–2721, 2019.

[8] Xu, H., Liu, B., Shu, L., and Yu, P. S. BERT post-training for review reading comprehension and aspect-based sentiment analysis. *Proceedings of EMNLP*, pp. 2324–2335, 2020.

[9] Jindal, N., and Liu, B. Opinion spam and analysis. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 219–230, 2008.

[10] Hu, Minqing and Liu, Bing. Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177, 2004.

[11] Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 188–197, 2019.

[12] Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., and Ma, S. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 83–92, 2014.

[13] 新井 智也，佐藤 哲司. 評価視点別の言及度を用いた意見文の分類手法の提案. 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), A2-2, 2010.

[14] 松波 友稀，上田 真由美，中島 伸介，階上 猛，岩崎 素直，O'Donovan John，Kang Byungkyu. コスメアイテム評価表現辞書を用いた評価項目別レビュー自動スコアリング方式. 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM2016), C2-3, 2016.

[15] 曽田 颯人，白井 清昭. 商品レビューの複数の観点からの有用性の評価. 言語処理学会第 27 回年次大会発表論文集,pp. 518–522, 2021.

[16] Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017.

[17] Meng, Y., Shen, J., Zhang, C., and Han, J. Weakly-Supervised Neural Text Classification. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 983–992, 2018.

[18] Wang, Y., Yao, Y., Tong, D. S., Xu, Z., and Drucker, S. M. Want to reduce labeling cost? GPT-3 can help. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4195–4205, 2021.

[19] Ding, B., Qin, C., Liu, L., Bing, L., Joty, S., and Li, B. Is GPT-3 a Good Data Annotator? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11173–11195, 2023.

[20] Gilardi, F., Alizadeh, M., and Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, e2305016120, 2023.

[21] Wang, Yizhong, Kordi, Yeganeh, Mishra, Swaroop, Liu, Alisa, Smith, Noah A., Khashabi, Daniel, and Hajishirzi, Hannaneh. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13484–13508, 2023.

[22] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Advances in Neural Information Processing Systems*, vol. 35, pp. 16344–16359, 2022.

[23] Press, Ofir, Smith, Noah A., and Lewis, Mike. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022.

[24] Liu, Nelson F., Lin, Kevin, Hewitt, John, Paranjape, Ashwin, Bevilacqua, Michele, Petroni, Fabio, and Liang, Percy. Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, Vol. 12, pp. 157–173, 2024.

[25] 日本語評価極性辞書, https://www.cl.ecei.tohoku.ac.jp/Open_Resources-Japanese_Sentiment_Polarity_Dictionary.html

[26] 楽天データセット, https://alaginrc.nict.go.jp/rakuten-dataset.html

[27] Tohoku University. Japanese BERT with Whole Word Masking. Available at: https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking.

[28] LLM-jp Project. Japanese ModernBERT. Available at: https://huggingface.co/llm-jp/llm-jp-modernbert-base.

[29] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[30] Liu, Y., Ott, M., Goyal, N., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[31] He, P., Liu, X., Gao, J., and Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *Proceedings of ICLR*, 2021.

[32] Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*, 2020.