レビューの多角的な有用性判別のための分析と分類モデルの構築

屋比久博文 [1]　當間愛晃 [2]
[1] 琉球大学大学院 理工学研究科 知能情報プログラム
[2] 琉球大学工学部工学科知能情報コース
k238571@ie.u-ryukyu.ac.jp　tnal@ie.u-ryukyu.ac.jp

概要

多角的な有用性判別の実現を目的として EC サイトのレビュー体系分析を行った．分析の結果，レビューの傾向は「日常的に利用する商品」と「一時的に利用する商品」で異なることがわかった．「日常的に利用する商品」に焦点をあてさらに分析することで 9 つのラベルを考案し，各ラベルに自動で分類するモデルの構築と精度向上のための 3 つのパターンで実験を行った．その結果，損失関数の重み調整が精度向上に寄与することが確認された．分類精度の更なる改善に向けて，オーギュメンテーション，不均衡データになりづらいラベル設計や交差検証の検討が必要であると考えている．

1　はじめに

インターネットの普及により商品を購入する際，SNS やレビューサイトを利用して商品情報を収集する人が増加している．さらにオンラインショッピングサイトである Amazon[1)] や楽天[2)] の登場によりその利用者と商品に対するレビューもまた増加傾向にある．しかし，膨大なレビューのすべてに目を通し，必要なレビューを収集することはユーザーにとって大きな負担となる．

この問題に対して，「Amazon レビュー文の有用性判別実験 [1]」や「有用なレビューを抽出するための比較文フィルタリングの検討 [2]」のようにレビュー文がユーザーにとって有用であるかを判別する研究は多く行われている．これらの研究では「購入するかどうかの意思決定に寄与する文を，ユーザーにとって有用な文」と位置付け，各ユーザーの趣味趣向を考慮しない，有用であるかどうかの 2 値分類を行っている．

また，Hong らは有用性であるかどうかの決定要因は，一貫性がなく有用性の測定法，レビュープラットホーム，製品タイプの 3 つの要因によって異なると主張している [3]．このように有用性の基準については多く議論されており，様々な観点から有用性を評価する研究なども行われている．

曽田らが行った「商品レビューの複数の観点からの有用性の評価 [4]」では有用性を「評価表現に対する根拠がある」，「商品に関係のある言及が多い」，「他の商品と比較している」，「実際に商品を使用した（あるいはしていない）と推測できる」，「評価（レーティング）に対する根拠がある」，「文量が多い」，「文章が読みやすい」といった 7 つの観点に分類し，前者 3 つの観点の評価を実現した．しかし，有用性の様々な視点からの評価という面においては実現できている 3 つの観点では不十分であると言える．例えば，「商品に関係のある言及が多い」という観点で商品配送についてのレビューは除外されているが，ユーザーによってはこのレビューを重要視するケースも考えられる．また，「関係のある言及」がどのようなものか明瞭であるとユーザーにとってさらに多角的な評価となり有益である．

本研究では，多角的な有用性判別の実現を目的に EC サイトのレビュー体系を分析し，先行研究 [3, 4] の知見や消費者庁「消費者意識基本調査 [5]」を踏まえ，レビューに付随するラベルの考案と付与を行った．また，考案したラベルを基に各カテゴリに自動で分類できるような自然言処理モデルの構築に取り組んだ．

2　実験設計

1）　https://www.amazon.com/
2）　https://www.rakuten.com/

表 1　それぞれのタイプのレビュー例

	レビュー例
ゲーム（ピクミン）	ピクミンは可愛いし内容も面白い。他の皆さんが言ってるようにプレイ時間が短いので星 4 つ。永遠に遊べるピクミン作って欲しいな。
電化製品（掃除機）	吸引力もしっかりあり、コード付きなので使いたい時に充電器にせず使えていいですが音は大きめです

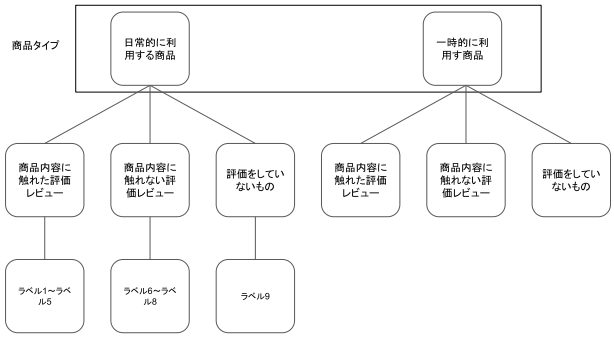

図 1　レビュー体系図

2.1　ラベル考案

Amazon カスタマーレビューよりレビューを収集し，商品タイプによるレビュー傾向の違いを観察した．その結果，電化製品や日用品など継続的に利用する商品と映画やゲームなど一時的に利用する商品との間に異なるレビュー傾向が見られた．表 1 にそれぞれのタイプのレビュー例を示す．

ゲームのレビューでは「面白い」や「楽しい」といった一時的な感情表現を用いて評価をする傾向が見られるのに対して，電化製品のレビューでは機能面での実用性を述べる日常的に利用することを前提とした評価をする傾向が見られた．この分析より本研究では商品タイプを「一時的に利用する商品」と「日常的に利用する商品」という 2 タイプに分けることとし，これらに付随するレビューの傾向が異なると仮定した．また上記 2 タイプの各商品には「商品内容に触れた評価レビュー」,「商品内容に触れない評価レビュー」,「評価をしていないもの」の 3 タイプのレビューが見られ，この 3 タイプのレビューをさらに細分化したレビューの評価基準をラベルとした．なお今回は「日常的に利用する商品」に着目してレビューを収集した．レビューの体系を図 1 に示す．

考案したラベルを以下に示す．

ラベル 1　商品機能に触れたレビュー
商品のサイズや性能面についての言及を含むレビューがこのラベルに該当する．
レビュー例：風量は思ったより強いですが、音も静かだし、プラズマクラスターだし気に入っています。

ラベル 2　別の商品内容のレビュー
レビュー元の商品から逸脱し，別の商品の解説を行なっているレビューがこのラベルに該当する．
レビュー例：最近買ったホームセンター製の扇風機が買って 2 年ほどで廻らなくなりました。やはり国内メーカーの製品が良いかな

ラベル 3　別の商品との比較を交えたレビュー
レビュー元の商品と別の商品の比較を交えて評価しているレビューがこのラベルに該当する．以下のレビュー例はラベル 1 と 3 の複合型となっている．
レビュー例：扇風機って長持ちするから、20 年前の物との比較ですが、静か、梱包がコンパクト、組立簡単でした。また 20 年持つのかな？

ラベル 4　商品価格に触れたレビュー
商品の価格表現（安い，高い，コスパ等）を含むレビューがこのラベルに該当する．
レビュー例：各部屋にコスパよし

ラベル 5　ユースケースを述べたレビュー
商品の活用例や体験談を交えて評価をしているレビューがこのラベルに該当する．以下のレビュー例はラベル 1 と 5 の複合型となっている．
レビュー例：静粛性は抜群で　熱帯夜の夜は　エアコンとの合わせ技で大活躍でした。表示関係も派手じゃないので　夜間まぶしいこともなく　良かったです。強いてあげれば　あまりにも重いのが難点で　掃除の時に場所をずらすのがちょっと気合いが筆世です。

ラベル 6　簡易的な感想表現のみのレビュー
商品内容に触れず，良し悪しのみで評価するレビューがこのラベルに該当する．
レビュー例：心地よい。

ラベル 7　商品状態に触れたレビュー

商品の配送時に生じる商品状態の変化や中古品の保存状態を評価したレビューがこのラベルに該当する．
レビュー例：傷がついていた事以外は機能面に問題はなし。しかし、外装もしょうひんではと、、、

ラベル 8　配送条件に触れたレビュー
商品配送期間や商品配送時のサービスを評価したレビューがこのラベルに該当する．
レビュー例：注文した翌日には商品が届きました。よい買い物をさせていただきました。

ラベル 9　批評をしていないもの
批評を行なっておらず，レビューとは言い難いものがこのラベルに該当する．
レビュー例：特にありません。

2.2　実験手順

本実験では BERT[6] を用いたファインチューニングによってマルチラベル分類モデルの構築を行った．行った 3 パターンの実験を以下に示す．

パターン 1　100 件のレビューに対し 9 つのラベルを付与しマルチラベル分類を行う．
パターン 2　766 件のレビューに対し 9 つのラベルを付与しマルチラベル分類を行う．
パターン 3　パターン 2 の分類をベースに全データの合計数と各ラベルの比率に応じた重みを加えたマルチラベル分類を行う．

実験手順の詳細を以下に示す．

- 1 つの商品から収集した 100 件のレビューをパターン 1 のデータ，さらに追加した 7 つの商品レビューを合計した 766 件のレビューをパターン 2 とパターン 3 のデータとした．
- 2.1 節で定義したラベルを基にアノテーションを行い，データセットを構築する．
- データセットを train : val : test = 6 : 2 : 2 の割合で分割する．学習率は 1e-6 とし，max エポックは 50 で 10 エポックのうち検証データに対する損失が改善されない場合，学習を終了する条件を加えた．
- 事前学習済み日本語 BERT モデル（‘cl-tohoku/bert-base-japanesewhole-word-masking’）を基にファインチューニングを行う．

表 2　100 件のレビューで学習した分類結果（パターン 1）

ラベル種別 (該当レビュー数/各ラベル毎の総数)	accuracy	recall	precision	f1score
ラベル 1 (15/68)	0.79	**1.00**	0.79	0.88
ラベル 2 (0/0)	**1.00**	0.00	0.00	0.00
ラベル 3 (1/2)	**0.95**	0.00	0.00	0.00
ラベル 4 (5/22)	0.68	0.00	0.00	0.00
ラベル 5 (13/52)	0.63	**0.92**	**0.67**	**0.78**
ラベル 6 (0/6)	0.95	0.00	0.00	0.00
ラベル 7 (0/6)	**0.95**	0.00	0.00	0.00
ラベル 8 (1/4)	0.89	0.00	0.00	0.00
ラベル 9 (0/0)	1.00	0.00	0.00	0.00

パターン 3 では上記に加えて以下の設定を行った．

- 全データ数に対する各ラベルの逆比に 0.4 を乗算したものを重みとして設定した．0.4 という数値は，パターン 2 の重みデフォルト値が 1 であるのに対して逆比が大きくなりすぎることから，スケールを縮小するため著者が直感的に設定したものである．式は以下のようになる．

$$\text{各ラベルの重み} = \frac{\text{全データ数 766 件}}{\text{各ラベル数}} \times 0.4 \tag{1}$$

2.3　評価手法

本研究では分類学習の評価手法として多く利用されている accuracy，recall，precision，f1 score を指標に分類結果の指標とした．

3　実験結果と考察

表 2 にパターン 1，表 3 にパターン 2，表 4 にパターン 3 の test データにおける分類結果を示す．それぞれの表は各ラベル（test ラベル数 / train，val，test ラベルの総数）という形式である．また各ラベル毎の accuracy，recall，precision，f1score における値が 0 以外のベストスコアは強調表示している．

3.1　パターン 1 とパターン 2 の比較

表 2 および表 3 に注目して観察すると，比較的全データ数における割合が大きいラベル 1 では 2 つのパターンで高い値を取っている．また，データ数の増加によってラベル 5 の accuracy，recall，precision，f1 score ともに数値が減少していることがわかる．この数値の低下は，パターン 1 においてラベル 5 が全データの約 50 %を占める一方で，パターン 2 で

表 3　766 件のレビューで学習した分類結果（パターン 2）

ラベル種別 (該当レビュー数/各ラベル毎の総数)	accuracy	recall	precision	f1score
ラベル 1 (136/660)	**0.88**	**1.00**	**0.88**	**0.94**
ラベル 2 (0/2)	**1.00**	0.00	0.00	0.00
ラベル 3 (27/88)	0.82	0.00	0.00	0.00
ラベル 4 (34/154)	0.76	0.00	0.00	0.00
ラベル 5 (62/231)	**0.60**	0.02	0.33	0.07
ラベル 6 (5/32)	**0.97**	0.00	0.00	0.00
ラベル 7 (8/20)	**0.95**	0.00	0.00	0.00
ラベル 8 (10/19)	0.94	0.00	0.00	0.00
ラベル 9 (3/18)	0.98	0.00	0.00	0.00

表 4　766 件のレビューと各ラベルに重みを加えて学習した分類結果（パターン 3）

ラベル種別 (該当レビュー数/各ラベル毎の総数)	accuracy	recall	precision	f1score
ラベル 1 (133/660)	0.86	**1.00**	0.86	0.92
ラベル 2 (0/2)	**1.00**	0.00	0.00	0.00
ラベル 3 (23/88)	0.76	**0.35**	**0.27**	**0.30**
ラベル 4 (33/154)	**0.79**	0.00	0.00	0.00
ラベル 5 (67/231)	0.49	0.73	0.45	0.56
ラベル 6 (8/32)	0.95	0.00	0.00	0.00
ラベル 7 (10/20)	0.92	0.00	0.00	0.00
ラベル 8 (7/19)	**0.95**	0.00	0.00	0.00
ラベル 9 (0/18)	**1.00**	0.00	0.00	0.00

はその割合が 30 %に減少していることに起因していると考えられる．

上述のことから分類モデルにおける各ラベルの損失を調整することや各ラベル数と全データ数との均衡を保つことが精度改善につながると言える．

3.2　パターン 2 とパターン 3 の比較

表 3 および表 4 に注目して観察すると，重みを調整したことによりラベル 3 とラベル 5 における評価の値に改善が見られる．このことから節 3.1 で述べた損失を調整することによりモデルの性能向上に寄与する可能性を裏付けるものとなった．

ラベル 2，6，7，8，9 では精度の改善は見られなかった．ラベル 2，7，8，9 は節 3.1 で述べたように，全データ数に対する各データ数が極端に少ないことが主な要因となっており，ラベルの重み調整による改善は見込めないと考えている．少数ラベルに対する重みをさらに増加させた場合でも，ほぼ完全一致のレビューのみ判定するようになると考えられるためである．

こういった極端に少ないデータ数に講じる策としては，純粋に偏ったデータを増加させるオーギュメンテーション，損失関数の設計を見直す，人の価値観 [7] を利用したラベル設計で不均衡データを避けることなどが考えられ，これらの手法について今後検討が必要であることがわかった．

ラベル 6 は少数ということの他に，val と test に割り当てられたデータ数が起因していることも考えられる．こちらは交差検証を行うことで再度検証．

また，ラベル 4 についてはラベル 3 と比較してデータ数が多いにも関わらず評価が改善できなかった．現状この問題について分析できておらず検討の必要がある．

4　まとめと今後の課題

Amazon レビュー文の分析を行い体系化を試みた．レビュー文は「一時的に利用する商品」と「継続的に利用する商品」の 2 つの商品タイプで傾向が異なることが確認できた．「日常的に利用する商品」の体系に基づいてレビュー文に対する分類ラベルを設計し，日本語 BERT モデルを用いてファインチューニングを行った．各ラベルの重み調整によりモデルの性能改善が一部見込めるが，極端に少ないデータセットに対しては有効でないことがわかった．今後は対策としてオーギュメンテーションや損失関数の再設計，不均衡データになりづらいラベル設計などの手法を検討していく必要がある．また，「一時的に利用する商品」のレビュー収集を進め分類モデルの構築に取り組んでいきたい．

参考文献

[1] 山澤美由紀, 吉村宏樹. Amazon レビュー文の有用性判別実験. 情報処理学会研究報告自然言語処理（NL）, 53(2006-NL-173) 号, pp. 15–20, 2006.

[2] 小橋賢介, 雨宮佑基, 酒井哲也. 有用なレビューを抽出するための比較文フィルタリングの検討. **DEIM Forum 2021 H11-4**, 2021.

[3] Hong Hong, Di Xu, G. AlanWang, WeiguoFan. Understanding the determinants of online review helpfulness: A meta-analytic investigation. **Support Systems**, Vol. 102, pp. 1–11, 2017.

[4] 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, 2021.

[5] 消費者庁. 【本文】令和 5 年版 消費者白書, 2024.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, Vol. abs/1810.04805, , 2018.

[7] Milton Rokeach. The nature of human values. **Contemporary Sociology**, Vol. 5, No. 1, pp. 13–16 (4 pages), 1976.