E-029

レビュアのパーソナリティを考慮したレビュー分類手法
Review Classification Methods that Take into Account Reviewer Personality

川上 大凱[1)]　鈴木 優[1)]
Taiga Kawakami　Yu Suzuki

1　はじめに

EC サイトに投稿されるレビューはマーケティングに活用するため，レビューの評価値予測やレビュアの感情推定などの機械学習タスクを利用した様々な分析がなされている．しかしレビューの中には，ポジティブな評価ラベルが付与されているにもかかわらず改善点のみが記述されている場合など，レビューと評価ラベルに差異がある場合がある．我々はこの差異がレビュアごとに存在し，評価傾向というパーソナリティを表していると考えた．そこで，我々は評価傾向をレビューから予測される評価ラベルと実際に付与されている評価ラベルの差の平均と定義する．予測された評価値と実際の評価値の差が小さいレビューを書くことが多いレビュアは，ポジティブな評価とレビューを書いているため，評価をレビューに反映させていると言える．反対にレビューから予測された評価値と実際の評価値が 1 と 5 のように差が大きいレビュアは，評価値がポジティブである一方でネガティブなレビューを書いているため，評価をレビューにあまり反映させていないと言える．このようなレビュアはレビューから推測できる感情と実際の評価値が異なるため，評価値予測が困難になり，精度が不十分となってしまう．この問題を解決するため，我々はレビュアの評価傾向を評価値予測に利用する．

本稿において，我々は評価傾向がレビューとその評価値に含まれると考え，レビューと評価値のペアをそのまま推論時の参考データとして入力する．この際，LLM による推論時に参考データとして適さないノイズデータが含まれてしまう可能性がある．この問題を解決するため，我々は LLM に提示する参考データを同一人物や類似人物のレビューと評価値のペアとする．

実験の結果，LLM を用いたレビューの評価値予測において，評価傾向を参考データとして入力することによって RMSELoss が 1.1 から 0.8 に減少し，精度が向上することがわかった．

2　関連研究

櫻井ら [1] は LLM を用いた著者識別に Prompt Tuning を用いることを提案した．実験結果より，指示文に Prompt Tuning を適用した場合，推論対象と異なる著者の文章を例示した場合であっても識別精度の低下がほとんど確認されなかった．本研究において，我々は Prompt Tuning よりも学習コストが低い Few-Shot Prompting を用いて推論対象と異なる著者の文章を例示する．

鈴木ら [2] はレビュアの執筆傾向から推測したスタイル（否定的・肯定的）やスタンス（客観的・感情的・ユーモア的）に基づいて類似レビューを推薦する手法を提案した．実験結果より，提案手法は文書ベクトルを参照したベースライン手法と比べて感性面におけるペアワイズ比較の勝率が有意に向上したが，文書内容におけるペアワイズ比較の勝率は有意に向上したと言えなかった．本研究において，我々はレビュアの執筆傾向に着目するのではなく，レビュアの評価傾向に着目する．

1）　岐阜大学大学院 自然科学技術研究科

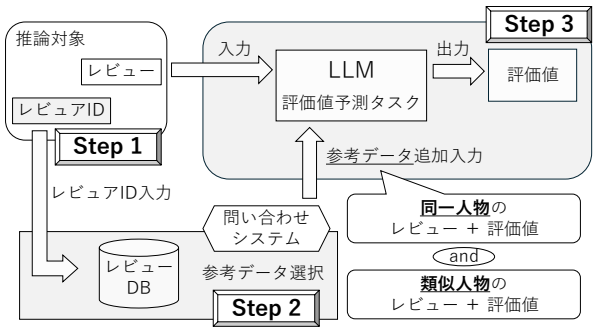

図 1　提案手法の概要図

3　提案手法

本研究の目的は，レビューを対象とした精度向上のために，レビュアの評価傾向に着目した参考データとしてふさわしいしいレビューとその評価値を選択することである．提案手法は図 1 に示す．

Step 1 レビュア ID を用意
Step 2 レビューデータベースを用いて参考データを選択
Step 3 レビューと参考データを LLM へ入力し評価値予測

Step 1 で我々は推論対象のレビューに付随しているレビュア ID を抽出する．Step 2 で我々はレビュア ID をレビューデータベースに問い合わせ，LLM の推論を補助する参考データの検索と選択を行う．Step 3 で我々は推論対象のレビューと参考データを入力とした LLM による評価値予測を行う．我々は Step 2 と Step 3 について 3.1 節と 3.2 節で説明する．

3.1　参考データの選択 (Step 2)

参考データを選択するため，我々はレビュアと同一人物のレビューと評価値を利用する方法と，追加で類似人物のレビューと評価値も利用する方法の二つを考えた．レビューに付与される評価値は，同じレビューでもレビュアの評価傾向によって変化する．つまり，同一人物が書いたレビューは同一の評価傾向を基に評価値が付与され，類似した人物の書いたレビューは類似した評価傾向を基に評価値が付与される．よって我々は 3.1.1 節と 3.1.2 節で説明する二つの方法が参考データの選択に有用だと考えた．

3.1.1　同一人物のレビューと評価値利用

同一人物のレビューを選択するため，我々はレビューに付随するレビュア ID を利用する．この際，レビュアによってはレビュー数が少ない場合があるため，我々は一人のレビュアが持つレビュー数に下限を設定し，下限を超えているレビュアのレビューのみを利用する．また反対にレビュー数が多い場合があるため，本稿において我々はレビューに対して文字数の下限を設定し，下限を超えているレビューのみを利用する．なお本研究において我々は，該当するレビュアのレビューが少ない場合，3.1.2 節の方法を追加で利用し，文字数の下限を超えるレビューが多い場合，ランダムサンプリングを実施しレビュー数を減らす．下限条件は 3.1.2 節で述べる方法においても同様とする．

あなたは評価値予測のエキスパートです.
以下のデータを参考に評価値を予測してください.
===参考データ===
レビュー：日曜日に宿泊しましたが，大変コスパが良かったです．評価値：5
===推論対象のレビュー===
レビュー：前日に低価格で予約できたのはここだけでした.

図 2　Few-Shot Prompting の一例

3.1.2　類似人物のレビューと評価値利用

我々は，レビューごとの評価傾向の分布である度数分布を用いて類似度を測定する．度数分布はレビュアが実際に付与した評価値と，レビューから LLM が予測した評価値の差を階級とする．度数分布を作成するため，我々は LLM を用いてレビューを入力とした Zero-Shot での評価値予測タスクを解く．その後，二人のレビュアの度数分布間の類似度を測定する．最後に，それぞれのレビュアの度数分布に対してカイ二乗検定を行い，有意に類似しているならば類似した評価傾向を持つレビュアのレビューとその評価値を参考データとして利用する．

また，我々はレビュアが評価値を付与する上で重視している観点がどこなのかを明確に評価傾向へ反映させることによって，より詳細に評価傾向の類似度を算出することが可能だと考えた．そこで，我々はすべての評価項目を網羅した評価値の代わりに，一部の評価項目のみが反映されたサブ評価項目の評価値を用いる方法も考えた．サブ評価項目の評価値とは，対象のサービスや商品に対する総合評価ではなく，設備や料理，品質，料金のようなカテゴリごとの評価値を指す．つまり，サブ評価項目の評価値はレビュアが評価値を付与する上で重視している観点がどこなのかを明確に反映している．この場合においても，総合評価値を用いる際と同様の方法で類似度の測定を行う．

3.2　LLM への入力と評価値予測 (Step 3)

我々は推論対象のレビューと参考データを，LLM の推論精度を向上させる文脈内学習手法であり，少数の例を推論対象のデータと同時にプロンプトとして入力する Few-Shot Prompting を用いて LLM へと入力する．それにより，LLM は提示された例と事前学習時に獲得したタスクや言語表現のパターンを照合するため，タスクに適した出力構造のマッピング規則を模倣することが可能となる．我々は評価値予測タスクや推論対象のレビュー，評価傾向に対する LLM の理解度を深めることを目標として図 2 のようなプロンプトを用いる．

4　実験

我々はレビュアの評価傾向が精度向上に寄与していることを確認するため，3.1.1 節で述べた方法を用いて実験を行った．LLM に対して同一人物の書いたレビューと評価値を例示した場合と，非同一人物レビュアの書いたレビューと評価値を例示した場合でそれぞれ評価値予測タスクを実施し，精度を比較した．

4.1　実験準備

本実験におけるタスクは評価値予測タスクであり，入力がレビューで出力が 1〜5 の整数値である．使用するデータセットは楽天トラベルデータセット[1)]である．使用する LLM は，Llama3[2)] と Gemma2[3)] である．量子化は行わない．我々は 1 タスク 20 件のレビューを実験に使用し，推論対象データ 1 件，参考データ 19 件として Few-Shot Prompting を行う．精度の評価指標は，LLM の予測値が実際の評価値から大きく外れている場合とそうでない場合を精度に反映させるために RMSELoss を用いる．

表 1　テストデータに対する RMSE Loss

	提案手法	比較手法
Llama3	0.801	1.156
Gemma2	0.801	1.088

4.2　実験手順

実験では以下のプロセスを適用した.

1. 20 文字以上のレビューを 20 個以上持つレビュアのレビューと評価値，レビュア ID を 20 件取得する.
2. 参考データを以下の手法で取得する.
   (a) レビュア ID を 1 件選択し，レビューと評価値 20 件をランダムサンプリング (提案手法)
   (b) レビュア ID が被らないよう，レビューと評価値 20 件をランダムサンプリング (比較手法)
3. 2a. と 2b. の両データに対して Leave-One-Out Cross-Validation で推論を行い，精度を比較する.

LLM は想定外の出力を行うことがあるため，評価値が正常に予測されるまで最大 20 回の再出力を行った．上限回数に到達したが正常な出力が得られなかった場合，結果集計時にそのレビューを除外する.

4.3　結果・考察

表 1 より，Llama3 では RMSELoss が 1.156 から 0.801 に減少し，Gemma2 では RMSELoss が 1.088 から 0.801 に減少したことから，両 LLM において提案手法が比較手法の精度を上回っていることがわかった．提案手法は一人のレビュアが持つ一貫した評価傾向が表現されている一方，比較手法は複数のレビュアが持つ複数の評価傾向が表現されている．以上より，我々はレビュアの評価傾向が評価値予測の精度向上に寄与していると考える.

5　おわりに

本研究では，LLM を用いたレビューの評価値予測において，評価傾向を参考データとして入力することにより精度が向上することを確認した．今後は，3.1.2 節で述べた手法を用いた実験を行い，提案手法と比較したいと考えている．また我々は，提案手法の組み合わせや参考データのサンプリング手法の改善を行う必要があると考えている.

謝辞

本研究の一部は JSPS 科研費（24K03044，23K28383）の助成を受けたものです．本研究では，NII IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用しました.

参考文献

[1] 櫻井航, 浅野雅人, 井元大輔, 本間正勝, 黒沢健至. 日本語 llm の prompt tuning による著者識別. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2024.

[2] 鈴木醇, 牛尼剛聡. 視聴後の鑑賞支援のためのスタイルとスタンスに基づいたレビュー推薦手法. In *DEIM forum*, 2025.

1) 楽天グループ株式会社 (2014): 楽天データセット. 国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi.org/10.32130/idr.2.0
2) https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
3) https://huggingface.co/rinna/gemma-2-baku-2b-it