修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/11/2022 (MM/DD/YYYY)

専攻名（専門分野） Department	経営システム 工学専攻	氏 名 Name	山下　皓太郎 Koutarou Yamashita	指 導 教 員 Advisor	後藤　正幸　　印 Seal
研究指導名 Research guidance	情報数理応用研究	学籍番号 Student ID number	CD 5221C041-4		
研究題目 Title	BERT の特徴量抽出に基づく製品レビュー分析モデルに関する研究 A Study on Product Review Analysis Model Based on BERT Feature Extraction				

1. はじめに

近年，EC サイト上などに一般消費者が閲覧可能な形で，商品に対するレビューが大量に投稿されている．これらのレビューデータには製品を購入，使用しているユーザの感想や要望が記述されており，他のユーザの購買行動に直接的な影響を与える．一方で，メーカーにとっては，レビューデータを分析することがユーザのニーズ把握のみならず，既存製品の問題点などを把握するために重要となっている．また，これらのレビューデータには投稿ユーザによる評価値も付与されており，ユーザの直接的な商品評価情報としての分析価値もあると考えられる．ユーザは一般的にレビュー内容に合致する評価値を付与すると考えられるが，中には商品に不満を抱いていながら建前として高い評価値をつけているユーザや，商品に満足していながら適当に中間の評価値をつけているユーザも存在し，レビューの内容と評価値は必ずしも一致しない．このような場合に評価値に着目して商品の評価を判断してしまうと，実際のユーザからの評価とは乖離した商品評価を下してしまう可能性がある．

そこで本研究では，自然言語処理の分野において近年有効性が指摘されている BERT(Bidirectional Encoder Representations from Transformers) [1] と感情分析の手法を導入することで，レビュー文に表れているユーザの感情内容を考慮した評価値を推定する手法の提案を行う．提案手法で得られる評価値はレビュー内容を正しく反映しており，建前ではないユーザの本心を推定した評価値を得ることが可能となる．具体的な事例として，ショッピングモール型 EC サイトAに出品する，大手メーカーB社の主要商品のレビューデータに提案手法を適用し，その有用性を示す．

2. 問題設定

本研究が対象事例とするメーカー B 社では，レビューそれぞれに付与されたユーザからの 5 段階の評価値（1 が最も評価が低く，5 が最も評価が高い）に基づいて，ユーザの各商品に対する評価を把握している．しかし，レビューデータの中には，ネガティブな内容をレビュー文として記述しているにも関わらず，最も高い評価値である 5 を付けているユーザや，逆にレビュー文ではポジティブな内容を記述しているにも関わらず，最も低い評価値である 1 を付けているユーザも一定数存在している．また，自身の感情に関係なく，適当に中間の評価値である 3 をつけているユーザもいる．評価値付与はレビュー文の記入よりも容易な作業であるため，必ずしも絶対的に正しくユーザの評価の程度を表した数値とはなっていない．また，ユーザによって評価値付与に対する考え方が異なるという場合も考えられる．例えば，あるユーザは「非常によかった」というレビュー文を記入して最高評価値である 5 を付与するが，同じ「非常によかった」というレビュー文を投稿した別のユーザは，その 1 つ下の評価値である 4 を付与しているといったことも散見される．これは比較的容易に最高評価値を付与するという評価が甘めのユーザと，自身の中で特別満足した商品でなければ最高評価値は付与しない厳しい評価値を与えるユーザが混在していることが原因であると考えられる．このようなレビュー文と評価値が乖離したレビューに対して，レビュー文の内容を加味した，実際のユーザの感想により即した評価値を付与することは，商品の強みや問題点を正確に把握する上で重要であり，メーカーにとって価値が高いと言える．

3. 準備

本研究では，BERT を用いてレビューを特徴ベクトルへ変換する．BERT は大量のテキストデータによる事前学習済みモデルである．分析対象のテキストデータを学習してモデル構築を行う手法ではないため，今回の研究対象であるレビューデータのような件数が少なく，長文や短文が混在したデータに対しても，各文章の特徴を捉えたベクトル変換を行うことができると考えられる．

3.1. BERT

BERT は，2018 年に Google が発表した自然言語処理のための Transformer[2] を用いた汎用言語表現モデルである．Wikipedia などから得られる大規模な生テキストデータのコーパスでモデルを事前学習していることが特徴である．事前学習によって単語や文書の埋め込み表現を獲得し，その後，ファインチューニングによって実際に解き

たいタスクに合わせて学習を行うことも可能である．これにより，自然言語処理の様々なタスクで高い精度を獲得している．

3.2. VADER

VADER[3] は，辞書ベースの感情分析手法の一種である．VADER ではまず，様々な言語資源からネガティブ，ポジティブ（以降，ネガポジ）に関連する 9000 語程度の単語の抽出を行う．次に，事前に感情分析のトレーニングを行った数名の専門家によって各単語に-4 から 4 までの感情値を付与する（-4 が最もネガティブ，4 が最もポジティブ）．その後，専門家たちの平均的な感情値が一定以下であった単語や，感情値の分散が大きい単語等を削除していき，最終的に 7500 語程度の感情語辞書の作成を行う．文章に対して VADER を適用する際は，文章内に感情語辞書の単語が存在していればその単語に対応した感情値を付与する．文章内で付与された感情値の総和を取ることで，文章全体の内容を考慮した感情分析を行うことが可能である．多くの感情分析のタスクにおいて，VADER は人間の実際の感情に即した感情分析を行うことが可能な手法である．

3.3. 関連研究

レビュー文に対する感情分析は既にいくつかの研究がなされている．Valdivia ら [4] は，旅行サイトのレストランやホテルに対するレビューデータに対して，VADER による感情分析を用いることで，評価値とユーザによって付けられた評価値の集約モデルを提案し，レビュー文とユーザによる評価値の両方を加味した新たな評価値を作成することで，それらの乖離に対処した．この手法では感情分析で得た感情値に一定の重み付けを行い，ユーザの評価値との平均を取ることで，新たな評価値の作成に成功している．

4. 提案手法

4.1. 概要

提案手法ではユーザによる評価値に対し，レビュー内容を反映させた新たな評価値の作成を行う．具体的には，VADER によってレビュー文それぞれに感情ラベルの付与を行うことで，レビュー内容の中のユーザの感情値を定量化する．その後，既存のユーザによる評価値にこのラベルを組み合わせ，新たに評価値を作成する．これにより，レビュー文と評価値で異なる評価をしているユーザのレビューに対しても，レビュー内容が正しく反映されることが期待される．しかし，単に辞書内の感情値に即してレビュー中全ての単語を考慮した感情ラベルを定義してしまうと，レビューデータの特性を無視した感情値が付与されてしまう可能性がある．商品レビューの種類（レストランのレビューやプリンタのレビューといった）によって頻繁に用いられる単語は異なり，これらの単語は特にその商品の良し悪しを判断する際に使用される可能性が高い．そのため，どの種類の商品レビューにも登場するような単語群よりも，対象の商品レビューでのみ高頻度に出現する単語群のほうが強く評価に影響していると考えられる．また非常に長文のレビューに対しても，信頼性に欠ける感情値が定義されてしまう可能性がある．そこで本研究ではまず，レビューデータ内の出現頻度の多い単語から，感情語辞書に存在する単語を複数抽出する．次に，抽出された単語が含まれるレビューに，その単語の有無を示すラベルを付与し，これを感情ラベルと定義する．その後，そのラベルの有無と，レビューを BERT によって変換した高次元特徴ベクトルを教師データとして，ラベル付与されていないレビューへのラベル付与を機械学習を用いて行う．単語の有無によって感情ラベルが付与されていないレビューにも，意味合いが似た内容のレビューであれば同種の感情が定義できる可能性が高い．そのため，この STEP で感情ラベルごとに BERT と分類器によってラベル付与を行うことで，真の意味でその感情ラベルに即したレビューに満遍なくラベル付与を行うことが可能である．最後に，感情ラベルをもとにレビューそれぞれに感情値を付与し，それら感情値とユーザ評価値との平均を取ることで，新たな評価値の作成を行う．次に具体的な手順を示す．

STEP1）VADER による感情ラベルの付与

VADER の辞書をもとに，レビューの内容に即した感情ラベルを付与することを考える．まず全体のレビューデータ件数を N とし，その中で出現頻度の高い単語から順に L 個，VADER の感情語辞書に含まれる単語を抽出する．その際，ポジティブと定義されている単語を L_{pos} 個，ネガティブと定義されている単語を L_{neg} 個とし，$L_{pos} = L_{neg}$ となるように抽出を行う（$L_{pos} + L_{neg} = L$）．次に，これら L 個の抽出単語集合に含まれる単語が存在するか否かを各レビュー $n(n = 1, 2, \cdots, N)$ に対して確認し，あるラベル $l(l = 1, 2, \cdots, L)$ が存在すればラベルの有無を表す変数 $y_l^n = 1$ とし，存在しなければ $y_l^n = 0$ とする．

STEP2）機械学習による感情ラベルの付与

レビューにより適切な感情ラベルを付与するマルチラベル問題に対し，機械学習モデルによって感情ラベルごとの二値分類器を構築することで，STEP1 ではラベル付与されなかったレビューにもラベルを付与することを考える．また，本研究では大規模文書データにより事前学習済みの BERT を用いることで，レビュー文を特徴ベクトルに変換することが可能である．

ここで，サイズ M のラベル付きレビューデータ集合 $\mathcal{D}^{\mathcal{L}}$ とサイズ $N - M$ のラベル未付与レビューデータ集合 $\mathcal{D}^{\mathcal{T}}$ があるとき，$\mathcal{D}^{\mathcal{L}} \cup \mathcal{D}^{\mathcal{T}}$ 中の n 番目のレビューの BERT によって得られた D 次元特徴ベクトルを $\boldsymbol{x}_n \in \mathcal{R}^D$ とする．このとき，$\mathcal{D}^{\mathcal{T}}$ に含まれるラベル未付与のレビューにラベル l を付与するためのアルゴリズムを以下に示す．

Step2-1）

ラベルデータ $\mathcal{D}^{\mathcal{L}}$ に含まれる $\boldsymbol{x}_n$ を説明変数，ラベル l の有無を表す y_l^n を目的変数として，各ラベル $(l = 1, 2, \cdots$

, L) に対し，それぞれ二値分類器を構築する．

Step2-2）

構築した分類器を用いて，$\mathcal{D}^{\mathcal{T}}$ に含まれるラベル未付与のレビューに対し，ラベル予測を行う．

Step2-3）

ラベル予測の結果，ラベルが付与されると判断されたレビュー n に対しては $y_l^n = 1$ へと変換を行う．

STEP3）レビューへの感情値の付与

付与された感情ラベルをもとに，各レビューに感情値を付与することを考える．まず，レビュー n に付与されているポジティブなラベルである L_{pos} の個数の総和を計算し，その値を s_{pos}^n とする．すなわち，n 番目のレビューに付与されているポジティブなラベルの総数を表す変数が s_{pos}^n である．同様にネガティブなラベルである L_{neg} の個数の総和も計算を行い，その値を s_{neg}^n とする．最後に，各レビューの感情値を $s_n = s_{pos}^n - s_{neg}^n$ と定義する．このポジティブラベルとネガティブラベルの差異を考えることにより，各レビュー内容中に反映されたユーザの感情を，感情値として定義することが可能である．

STEP4）新たな評価値の作成

STEP3 で得られた感情値をもとに，よりユーザの本心に即していると考えられる新たな評価値の計算を行う．

まず，各レビューに付与されたユーザによる評価値を u_n として，n 番目のレビューにおける正規化後の感情値を式 (1)，正規化後のユーザ評価値を式 (2) で定義する．

$$s_n^c = \frac{s_n - s_{min}}{s_{max} - s_{min}} \tag{1}$$

$$u_n^c = \frac{u_n - u_{min}}{u_{max} - u_{min}} \tag{2}$$

ここで，式中の添字の min, max は，それぞれ $1 \sim N$ 番目までの s, u における，最小値と最大値であることを示す．この正規化のステップにより，各値が 0 から 1 までの連続値に変換される．その後，正規化後の 2 つの値に対し，次の式 (3)，式 (4) によって幾何平均，算術平均を計算する．幾何平均は，Valdivia らの研究で用いられている方法であるが，本研究では平均値の取り方による差異を検証するため，算術平均も加えた 2 種類の評価値を提案し，各評価値について推定精度を検証する．

$$u_n^g = (u_n^c)^{1-\beta} \cdot (s_n^c)^{\beta} \tag{3}$$

$$u_n^a = (1-\beta) \cdot (u_n^c) + \beta \cdot s_n^c \tag{4}$$

ここで，β は感情値を考慮する度合いを表すハイパーパラメータである．すなわち，このハイパーパラメータの値により，レビュー内容をどの程度考慮するか制御することができる．この式 (3)，式 (4) によって定義される u_n^g, u_n^a が今回の手法で提案する新たな評価値である．

表 1: ポジティブラベルの分類器の AUC

easy	great	love	happy	recommend
0.866	0.812	0.842	0.897	0.948

5. 実データ分析

5.1. 分析データおよび分析条件

メーカー B 社の主力商品であるレーザープリンタに対するサイト A 上の英文レビューデータに提案手法を適用する．データ期間は 2017 年 1 月〜2021 年 7 月，データ件数は $N = 13,953$ 件，BERT の事前学習モデルより獲得する特徴ベクトルの次元数は $D = 768$，用いる感情ラベルの数 $L = 10$ とし，感情ラベル名は，ポジティブなラベルに対しては easy，great，love，happy，recommend の 5 種類，ネガティブなラベルに対しては problem，hard，waste，bad，trouble の 5 種類とした（$L_{pos} = L_{neg} = 5$）．感情ラベルが付与されるレビュー数を拡張する際の，二値分類の評価指標には Area Under the ROC Curve(AUC) を用い，$\mathcal{D}^{\mathcal{L}}$ のうち 80% を学習，20% をテストデータとして評価した．STEP2 の感情ラベルの付与には，RBF カーネルを用いたサポートベクトルマシン（SVM）を採用した[1]．また，作成した新たな評価値の妥当性を検証するために，事前にレビュー内容を人手によって確認し，その内容を考慮して手動で付与した評価値との誤差の比較を行なった．ここでは，人手でレビューを確認し，評価値を付与する一連の作業をアノテーションと呼ぶことにする．アノテーションはランダムに抽出した 100 件のレビューに対して，ユーザ評価値と同様に 1 から 5 までの 5 段階で評価を行い，その後正規化を行なった．正規化後のアノテーション評価値に対し，ユーザによる評価値，感情値，幾何平均による新たな評価値，算術平均による新たな評価値の 4 つの値との誤差を比較した．なお，誤差の評価指標には平均二乗誤差（Mean Squared Error，MSE）を用いた．感情値を考慮する度合いであるハイパーパラメータ β の値は 0.0 から 1.0 の間で 0.10 間隔で変化させ，分析を行なった．

5.2. 分析結果

まず，感情ラベルそれぞれにおける分類器のテストデータに対する AUC を表 1，表 2 に示す．いずれのラベル分類器においても AUC の値は高く，BERT と二値分類器を用いることで，単語の完全一致のみによる方法ではラベル付与されていないレビューに対しても概ね正しくラベル付与ができていることが分かる．

次に，これらの分類器を用いて感情ラベルが付与されるレビュー数を拡張し，拡張後のレビューデータに対して新たな評価値の作成を行なった．アノテーションにより得られた評価値に対し，ユーザ評価値，感情値，幾何平均によ

[1] ロジスティック回帰・SVM・LightGBM の 3 手法で対象データへの精度比較を行い，最も精度の高かった SVM を本研究では採用した．

表 2: ネガティブラベルの分類器の AUC

problem	hard	waste	bad	trouble
0.820	0.767	0.938	0.800	0.785

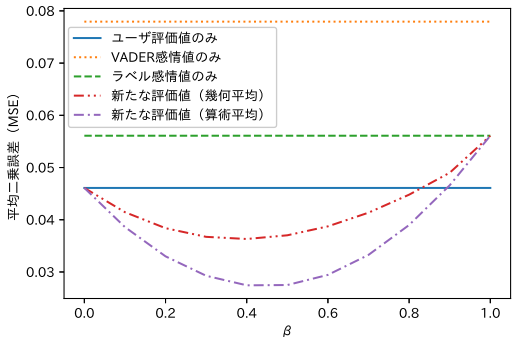

図 1: 各評価値の MSE（平均二乗誤差）

る新たな評価値，算術平均による新たな評価値の 4 つの値との誤差を比較した結果を図 1 に示す．

図 1 より，新たな評価値では適切な β の値を用いることによって幾何平均，算術平均のいずれにおいてもユーザ評価値のみ，および感情値のみの場合よりもアノーテーションした評価値との誤差が小さくできることが分かる．このことから 2 種類の平均値によって作成した新たな評価値はレビューの内容により即した評価値となっており，ユーザのレビュー文書中での評価と，ユーザが付与した評価値による評価との乖離を是正可能な指標であると考えられる．また最も誤差が小さくなった β の値は幾何平均と算術平均の場合でともに 0.40 となっており，適切な割合で感情値を考慮する必要がある可能性が示唆される．

6. 考察

アノテーションした評価値との誤差が最も小さかった算術平均による評価値を用いて，実際に全レビューに新たな評価値の付与を行なった．ただし，β の値は最も MSE が低かった 0.40 とした．また，レビューそれぞれにつけられたレビュー内容のカテゴリごとに，ユーザ評価値と新たな評価値の平均値を計算することで，レビューカテゴリごとの評価値の可視化を行なった．ここでカテゴリとは，レビューそれぞれの内容に基づいて付与された，そのレビューがどのような話題について話しているかを表す情報である（例：トナーについての内容であれば toner，価格についての内容であれば price のカテゴリとなる）．いずれの評価値も 0 から 1 までの連続値で表される．結果を表 3 に示す．

表 3 より，カテゴリごとに評価値は大きく異なることが分かる．speed，mobile，setup，quality，price，function については平均的な評価値が高く，商品に対してこれらのカテゴリの観点ではユーザの満足度が高い傾向がある．一方，paparfeed，service については平均的な評価値が低く，これらのカテゴリの観点では不満を抱えているユーザが一定

表 3: カテゴリごとの評価値の平均

カテゴリ名	ユーザ評価値	新たな評価値
speed	0.860	0.759
mobile	0.839	0.743
setup	0.817	0.735
quality	0.815	0.735
price	0.796	0.718
function	0.789	0.710
toner	0.724	0.660
network	0.677	0.623
paperfeed	0.631	0.588
service	0.444	0.446

数存在する可能性が高い．特に，service のカテゴリにおいては著しく評価値が低く，今後のマーケティング活動においては，各種サービス面においてメーカーによる対応の改善が必要であると考えられる．また，ユーザ評価値と新たな評価値を比較すると，全体として新たな評価値が低い傾向が見られる．このことより，ユーザの評価値はレビュー内容での評価よりも高い傾向があるといえる．

7. まとめと今後の課題

本研究では，VADER を用いた感情分析の結果をもとに，各レビューに感情値を付与し，その感情値とユーザから付与された評価値との 2 種類の平均値を計算することで，よりユーザのレビュー内容を反映した評価値を推定する手法を提案した．そして，実データを用いた分析により，提案手法がユーザによる評価値よりも，レビュー内容を反映した評価値を付与可能であることを示した．これにより，建前ではないユーザの真意に沿う商品評価を把握することが可能になると考えられる．

謝辞

本研究を行うにあたり，B 社から提供のレビューデータを使用させていただきました．貴重なデータの提供に深く感謝致します．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.

[3] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8, pp. 216–225, 2014.

[4] Ana Valdivia, Emiliya Hrabova, Iti Chaturvedi, M Victoria Luzón, Luigi Troiano, Erik Cambria, and Francisco Herrera. Inconsistencies on tripadvisor reviews: A unified index between users and sentiment analysis methods. *Neurocomputing*, Vol. 353, pp. 3–16, 2019.