

ホテルの評価値予測のためのカテゴリ特化型感情極性辞書

柴田 諒人¹ 亀井 清華^{2,a)} 中野 浩嗣^{2,b)}

受付日 2020年12月9日, 採録日 2021年3月31日

概要: ホテル予約サイトでは, ホテルの宿泊客から投稿された各ホテルに対する評価が提供される. 各評価には, 自然言語テキストによるレビューと, 「サービス」や「立地」, 「部屋」などの複数のカテゴリのそれぞれに対する数値による評価値が含まれる. しかし評価の中には, 一部の評価値がレビューの内容と矛盾しているものが存在する. そのような評価は, サイト利用者を混乱させる可能性がある. このような問題を解決するために, レビューの文章から各カテゴリに対する評価値を予測する手法を提案する. 本提案手法では, 楽天トラベルのレビューデータベースを利用して, カテゴリ特化型感情極性辞書を構築し, それを元に評価値を予測する. カテゴリ特化型感情極性辞書は, 「悪い」, 「軽い」, 「美味しい」などの各感情語について, カテゴリごとに感情極性値 (ポジティブまたはネガティブの度合い) を与える. 我々の実験では, BERT を用いて各レビューをベクトル化したレビューベクトルとカテゴリ特化型感情極性値を組み合わせることで, BERT を単体で用いる場合や既存辞書の感情極性値を組み合わせる場合よりも高い精度で評価値予測を行うことができた. さらに, カテゴリを細分化したサブカテゴリに対しても, カテゴリ特化型感情極性値と BERT とを組み合わせることの有用性を示した.

キーワード: 評価値予測, 自然言語処理, 感情極性, BERT

Category-oriented Japanese Sentiment Polarity Dictionary for Rating Prediction of Hotels

AKITO SHIBATA¹ SAYAKA KAMEI^{2,a)} KOJI NAKANO^{2,b)}

Received: December 9, 2020, Accepted: March 31, 2021

Abstract: Hotel booking sites provide us hotel evaluations, each of which includes a textual review and numeric ratings for multiple categories such as “Service”, “Location”, “Rooms”, etc., submitted by a hotel guest. However, textual reviews and numeric ratings of some evaluations are inconsistent. For example, a textual review which praises the hotel is submitted with low ratings. Such evaluations may confuse site users. To resolve such problems, we propose a high accuracy method to predict a numeric rating for each category from a textual review. Our new idea is to use *Category-oriented Sentiment Polarity Dictionaries* (CSPD), each of which is automatically compiled for each category using a Rakuten Travel review database in advance. The CSPD gives the *sentiment polarity value* (i.e., the positivity/negativity value) for each sentiment word such as “bad”, “light”, and “delicious” for each category. In our experiments, we combine the sentiment polarity value by CSPD and word vectors derived by BERT to obtain an expected rating value from a textual review. Our experimental results show that the combined method with CSPD attains higher accuracy than the case using only BERT vectors and the case using existing dictionaries. Additionally, when we apply CSPD and BERT to sub-categories obtained by segmenting each category further, our method is effective against the rating prediction from reviews.

Keywords: rating prediction, natural language processing, sentiment analysis, BERT

¹ ジェイアール東海情報システム株式会社
JR TOKAI Information Systems Company, Nagoya, Aichi
461-0022, Japan

² 広島大学大学院先進理工系科学研究科,
Graduate School of Advanced Science and Engineering,
Hiroshima University, Higashi-hiroshima, Hiroshima 739-
8521, Japan

1. はじめに

楽天トラベル [2] や TripAdvisor [3] などのホテル予約サ

^{a)} s10kamei@hiroshima-u.ac.jp

^{b)} nakano@cs.hiroshima-u.ac.jp

本稿は, 国際会議 WANC2020 [1] で発表した論文の拡張版です.

表 1 評価値とレビューの内容が不一致している例

Table 1 Examples of the mismatches between the review and the ratings.

	レビュー	評価値					
		サービス	立地	部屋	設備	風呂	食事
1	24 時間受付していたので便利でした。駅から近くて、コンビニも近くにあり便利でした。朝食時混雑していて落ち着いて食べることができなかった。	3	2	3	3	2	2
2	ロビーにあります機械のコーヒーと朝食のコーヒーが美味しくないです。コーヒー好きの私は長い長い 3 日間でした。次回はコーヒーが美味しいことを祈ります。	5	5	5	5	5	5

イトでは、各ホテルの利用客（レビュアー）から投稿された、各ホテルに対する評価が提供される。各評価は、自然言語によってホテルの印象を記述したレビューと、満足度を表す数値による評価値の組として投稿される。サイトの閲覧者にとって評価値は、自然言語で書かれたレビューよりも直感的であるといえる。その結果、閲覧者はレビューを考慮せず、評価値のみに頼ってホテルの情報を得ることが多い。しかし、各レビュアーはそれぞれ独自の評価の付け方をしており、同じホテルに対する、または同じレビュアーによるものであるにもかかわらず、いくつかのレビューと評価値の組は、サイトの閲覧者に対して異なる印象を与え、閲覧者を混乱させる可能性がある。たとえば、実際にあったレビューと評価値の組として、表 1 がある。1 番のレビューの場合、レビューの内容としてはそのホテルを褒めているにもかかわらず、それに付加された評価値が低いため、実際の評価が伝わらない。2 番のレビューの場合はこの逆の現象が起きている。これは、レビューに評価値の理由が書かれていないなどの記述不足か、あるいは評価値が誤って付加されたことが原因と考えられる。

そこで本研究では、レビューから評価値を予測することを考える。それにより、レビュアーがレビューを書く際に、レビュアーの評価値がレビューの内容と極端に離れている場合、レビュアーのレビューから予測される評価値に基づいて、レビュアーに確認を促すことができるものと考えられる。これにより、レビュアーは自分の評価値やレビューの内容を再考することになり、結果的に誤って選択された評価値や記述不足のレビューの数を減らすことができる可能性がある。これは、サイトの利用者だけでなく、ホテルオーナーにとっても宿泊者の意見を知るうえで有用であると考えられる。

本稿では、楽天トラベルのホテル予約サイトにおけるレビューからの各カテゴリに対する評価値の予測に焦点を当てている。楽天トラベルサイトにおいて、レビュアーは宿泊したホテルに対して以下の要素を持つ評価を投稿する。

- レビュー：文章による評価。
- 評価値：数値による評価。サービス、立地、部屋、設備・アメニティ（以下では設備と呼ぶ）、風呂、食事の 6 つの各カテゴリについて、各宿泊客が 1～5 の評価値を付けたり、評価値を付けなかったりすることが可能

となっている。

文章の単語を数値として特徴付けるための手法として単語のベクトル化手法がある。単語ベクトルとは、各単語を数百～数千次元の実数値ベクトルとして表すものである。単語間の意味的關係を、cos 類似度やユークリッド距離などを用いて表すことができる。提案手法では、レビューから各カテゴリを説明する文を抽出し、それらの各単語をベクトル化するための手法として、BERT [4] を使用した。作成した単語ベクトルの平均を用いて文を表すベクトルとする。我々と同じカテゴリ別評価値予測の問題設定について、外山ら [5]、張ら [6]、Lamdahbin ら [7] もベクトル化手法を用いて評価値を予測した。外山らは PV-DM [8] を、張らは Bag-of-Words を、Lamdahbin らは Word2Vec [9] または fastText [10] を使用し、レビューまたは単語をベクトル化して用いた。BERT は、その汎用性と精度の高さから現在最も注目を集めている単語ベクトル化手法の 1 つである。

レビューに基づく評価値予測にとって、レビューに表れる感情を考慮することは非常に重要だと考えられる。しかし、上記のベクトル化手法では、各単語が表すポジティブな感情とネガティブな感情を把握することは困難である。文中の各単語のポジティブまたはネガティブの度合いを感情極性値と呼び、感情極性値を持つ単語は感情極性語と呼ぶ。一般に、感情極性値が高い感情極性語はよりポジティブな意味を表し、感情極性値が低い感情極性語はよりネガティブな意味を表す。このような極性を取得するために、感情極性語と感情極性値の組の集合から構成される、いくつかの感情極性辞書 [11], [12], [13] が知られている。

文中の単語を感情極性辞書と比較することにより、その文が賞賛または批判をどの程度表しているかを示すことができる。しかしレビューでは、記述されているカテゴリによって各単語の意味合いが異なるものと考えられ、一般的な感情極性辞書の感情極性値はレビューには適さない [14]。そこで我々は、各感情極性語にカテゴリごとに固有の実数の感情極性値を持たせることで、評価値の予測精度を向上させることができると考えた。本稿では、カテゴリ特化型感情極性辞書（CSPD：Category-oriented Sentiment Polarity Dictionary）の作成手法を提案する。CSPD の例を表 2 に示す。CSPD では、レビューデータセットにおける、各カテゴリと各評価値の単語の出現率に基づいて、

表 2 カテゴリ特化型感情極性辞書の例

Table 2 Example of CSPD.

	サービス	立地	部屋	設備	風呂	食事
悪い	-5.7	-6.5	-4.6	-4.8	-5.8	-5.1
軽い	-2.2	1.3	-0.5	0.6	-0.6	0.05
美味しい	3.4	極性値なし	3.4	3.2	3.7	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

各単語のカテゴリごとの感情極性値を計算する。

本稿では、CSPD を単語のベクトル化手法である BERT と組み合わせて用いることの有効性も示す。さらに、カテゴリを細分化させたサブカテゴリを定義し、CSPD と BERT を用いて、サブカテゴリごとの評価値の予測も行い、有効性を示す。

本稿の構成は以下のとおりである。2 章では、関連研究について説明する。3 章では、提案手法について説明し、4 章で実験の評価について説明する。最後に 5 章で本研究のまとめと今後の課題について述べる。

2. 関連研究

2.1 感情極性辞書

まず、既存の英語の感情極性辞書について紹介する。

Thelwall ら [11] は、感情極性辞書を作成するために、文のポジティブとネガティブの値を整数で測定するシステムを作成した。彼らのシステムでは、事前に感情極性語と感情極性値を手作業で決定し、機械学習を用いてそれらを調整した後、いくつかのルールを使用して各単語の感情極性値を計算する。彼らの手法は、英語の感情分析の分野では非常に広く使用されているが、レビューに特化した単語の使われ方は考慮されていない。

レビューを対象としたものとして、文献 [15], [16], [17], [18], [19] などがある。Hu ら [15] は、30 の形容詞にポジティブまたはネガティブの 2 値の極性を人手で割り当て、それを基に WordNet を用いて同義語に同じ極性を、反意語に逆の極性を付けることで極性辞書を作成している。Popescu ら [16] の手法は、Hu らと同様に WordNet を用いて極性付けを行うが、さらに文の係り受けや接続詞から単語間の関係を決定する。そして、各単語の各極性の確率を学習することでポジティブ、ネガティブ、ニュートラルの 3 値の極性辞書を作成している。しかし、これらの手法では、各単語が使われたカテゴリやレビュー評価値については考慮されていない。Fang ら [17] は、各単語の評価値ごとの出現回数の加重平均を基に各単語の実数の極性値を決定する手法を提案している。この手法では、レビューで言及されているカテゴリは考慮されておらず、さらに、まだ出現回数が少ない単語にも大きな感情極性値が割り当てられてしまう場合がある。レビューのカテゴリごとに極性値を求める方法として、Lu ら [18] は、WordNet や既存の感

情極性辞書とレビュー全体への評価値など複数の根拠から得られる感情極性を組み合わせてカテゴリごとに 2 値の極性値を決定することを、最適化問題として定式化するアプローチを提案している。しかし、彼らの手法では文書全体での評価値がそのまま出現するすべての単語に反映されるため、1 つの文書の中に複数のカテゴリについて様々な感情が表現されるレビューでは望ましくない。Almatarneh ら [19] の手法では、カテゴリごとに各単語がポジティブな評価値のついたレビューとネガティブな評価値のついたレビューのどちらで出現率が高いかで、2 値の極性値を決定している。しかし彼らの手法では、極性の強弱が考慮されていない。

次に、日本語の自然言語処理の分野において広く用いられている 2 つの代表的な感情極性辞書を紹介する。1 つ目は、Takamura らによる単語感情極性対応表 [13] である。彼らは、各電子が持つスピンと呼ばれる方向（上向きか下向きか）と同じように、感情極性は各単語が持つ方向であると考え、単語をノードとする語彙ネットワークを構築し、平均場近似を利用して単語の感情極性値を判定した。この辞書の感情極性値は $-1 \sim 1$ の範囲の実数値である。もう 1 つは、乾ら [12] の日本語評価極性辞書である。この辞書の基となる辞書は小林ら [20] の評価値表現辞書であり、乾らは人手で 2 値の評価極性情報を付与した。しかし、これらの辞書はカテゴリごとの単語の使われ方の違いが考慮されていない。

本研究では、レビューのカテゴリごとの単語の使われ方に着目し、1 つのレビューの中で複数のカテゴリに対して様々な感情が表現されている場合でも有用な感情極性辞書の作成を考える。また、評価値予測に用いるために、単語間の感情極性値に差が出るようにし、さらに出現回数が小さい単語の影響度が強くないようにする。

2.2 単語ベクトル化手法

次に、本研究で用いる単語のベクトル化手法について紹介する。

Bojanowski ら [10] は、Word2Vec [9] の発展型である fastText を提案した。fastText は、単語をサブワードに分割することで活用形に対応する。例として、「go」と「goes」は Word2Vec では完全に異なる単語として解釈されたが、fastText では同じ「go」を含む単語として解釈される。fastText は単語のベクトルが、文脈に依存せずに単語ごとに独立して生成される。また、各単語を 100 次元程度の小さい次元のベクトルとしても表現が可能なので、単語どうしの類似度計算などの際に短時間で計算することができる。

BERT (Bidirectional Encoder Representations from Transformers) [4] は、文脈に沿った単語ベクトル化手法の 1 つであり、Devlin らによって提案された。双方向学習の MLM (Masked Language Model) と、次の文を予測し

ながら学習する NSP (Next Sentence Prediction) の 2 つの事前学習を行う。作成した BERT モデルを fine-tuning アプローチで、分類問題や意味読解など様々な問題にあったように調整することができる。

3. 提案手法

本研究では、レビューデータセット \mathbf{R} として楽天トラベルデータセット [21] を使用し、ホテルレビューのカテゴリ別評価値予測について議論する。

提案手法では、まず CSPD を作成する。データセット \mathbf{R} の各レビューを文ごとに区切り、各文で使用されている単語から、その文に記述されているカテゴリのラベルを文に割り当てる。次に、レビューごとに、各カテゴリを説明する部分のみを抜粋し、各カテゴリの各評価値に分類する (3.2 節)。そして、分類されたレビューと各単語の出現率に基づいて、カテゴリごとに感情極性値を計算し、CSPD を作成する (3.3 節)。

提案手法では、作成した CSPD を使用して、カテゴリに分類された文の単語に感情極性値を割り当てることで文の特徴量を計算し、評価値を予測する (3.4 節)。さらに、同じ CSPD を用いて、カテゴリをより細分化したサブカテゴリに対しても評価値予測を行う (3.5 節)。

3.1 カテゴリ辞書

まず、文に記述されているカテゴリのラベルをその文に割り当てるために、各カテゴリに関連する単語をまとめたカテゴリ辞書を作成する。カテゴリ辞書の要素の例を表 3 に示す。

カテゴリ辞書は、田熊ら [22] が作成した辞書に基づいている。これは、筑波大学文単位評価極性タグ付きコーパス [23] で使用されている上位 500 の名詞を手でカテゴリに分類したものである。彼らの辞書から、単語がレビューに含まれている場合にそのレビューに記述されているカテゴリを特定できるような 315 語のみを手で抽出した。このとき、カテゴリ辞書の各単語は、最大 3 つのカテゴリに属することができるものとした。

さらに、「美味しい」や「不味い」など、特定のカテゴリを明確に表す単語を追加するために、レビューデータセット \mathbf{R} 内で 10 回以上出現する各単語を fastText [10] を適用して 100 次元の単語ベクトルにする。このとき、各レビューを形態素解析エンジン Janome [24] を用いて単語ごとに分割し、すべての単語を基本形に変換する。そうして得られた単語ベクトルを用いて、上記の 315 語と他の単語間の cos 類似度を計算した。そして、cos 類似度が 0.8 以上である単語から、記述されたカテゴリを認識可能である単語を手で抽出し、追加した。

後述する旧データ (4.1 節) を \mathbf{R} として用いた場合におけるカテゴリ辞書の単語の内訳を表 4 に示す。

表 3 カテゴリ辞書の例

Table 3 Example of category dictionary.

サービス	立地	部屋	設備	風呂	食事
愛想	駅	TV	ベッド	温泉	デザート
価格	観光	ベッド	階段	浴槽	魚
⋮	⋮	⋮	⋮	⋮	⋮

表 4 カテゴリ辞書の登録語彙数の内訳

Table 4 Breakdown of the words in category dictionary.

	田熊ら	追加語	総語彙数
サービス	74	0	74
立地	50	0	50
部屋	96	1	97
設備	85	4	89
風呂	30	0	30
食事	46	10	56
総語彙数	315	15	330

3.2 カテゴリレビューの分類

次に、データセット \mathbf{R} の各レビュー r を「。」「.」「!」, 「?」など、文の終端を表す記号に基づいて文に分割する。そして、3.1 節で作成したカテゴリ辞書を用いて、分割した文をカテゴリに分類する。各文には、カテゴリ辞書の単語と照合され、カテゴリラベルが付けられる。複数のカテゴリの単語が含まれている場合、各文は複数のカテゴリに割り当てることができる。

各レビュー $r \in \mathbf{R}$ から $X \in \{\text{サービス, 立地, 部屋, 設備, 風呂, 食事}\}$ のカテゴリラベルが付加された文を抽出し、 X に関する文の集合 r_X をレビュー r ごとに作成する。この r_X を X のカテゴリレビューと呼び、 r を r_X の上位レビューと呼ぶ。また、 \mathbf{R}_X をすべてのレビューから抽出した X のカテゴリレビューの集合とする。

次に、各カテゴリレビュー r_X を X に対する評価値 $Y \in \{1, 2, 3, 4, 5, *\}$ ごとに分類する (ただし $*$ は評価値がない場合を表す)。つまり、 r に付属している X に対する評価値が Y であった場合に、 r_X をカテゴリレビュー集合 \mathbf{R}_{XY} に分類する。ただし、 \mathbf{R}_{X*} に分類された場合は、以下では使用しない。カテゴリレビューの分類の例を図 1 に示す。

3.3 感情極性辞書の作成

次に、カテゴリレビューから CSPD を作成する。感情極性語に使用する単語は、「形容詞」、「形容動詞」、「副詞」、「名詞 (形容動詞語幹)」の 4 つの品詞とする。データセット \mathbf{R} において、いずれかのカテゴリ X の \mathbf{R}_X に含まれるこれら 4 つの品詞の単語は感情極性語となる。CSPD では表 2 のように、各感情極性語に 6 次元のベクトルが与えられ、各要素として各カテゴリにおける実数の感情極性値が割り当てられる。感情極性語 A が \mathbf{R}_X に出現しない場合

には、 A のカテゴリ X の感情極性値は「極性値なし」とする。

各感情極性語 A のカテゴリ X における感情極性値 E_{XA} を計算する方法は、 \mathbf{R}_{XY} における A の出現率に基づいている。 \mathbf{R}_{XY} のカテゴリレビューの総数を $|\mathbf{R}_{XY}|$ とし、感情極性語 A を含む \mathbf{R}_{XY} のカテゴリレビューの総数を $|\mathbf{R}_{XYA}|$ とする。 \mathbf{R}_{XY} における A の出現率 U_{XYA} は、次のように計算される。

$$U_{XYA} = \frac{|\mathbf{R}_{XYA}|}{|\mathbf{R}_{XY}|}.$$

以下、感情極性値 E_{XA} の計算方法の詳細を説明する。

- (1) まず、下記の式を用いて、 X におけるすべての評価値での感情極性語 A の出現率 U_{XA} を計算する。

$$U_{XA} = \frac{\sum_{Y=1}^5 |\mathbf{R}_{XYA}|}{\sum_{Y=1}^5 |\mathbf{R}_{XY}|}.$$

- (2) \mathbf{R}_X 内のすべての感情極性語について X での出現率を計算した後、下記の式でそれらの平均値 U_{Xave} を計算する。

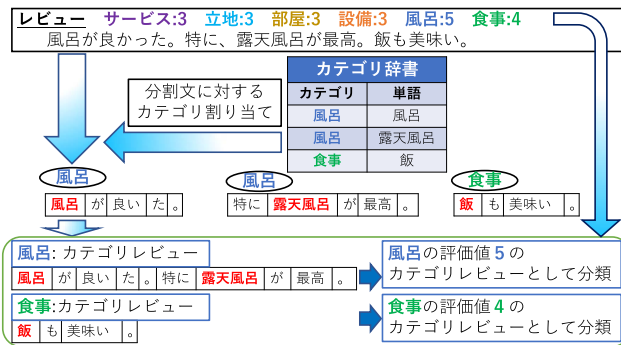


図 1 カテゴリレビューの分類の例

Fig. 1 Image of the category review classification.

表 5 CSPD の語彙数、基礎出現率、感情極性スコアの平均と標準偏差、および最大/最小の極性値を持つ単語

Table 5 Vocabulary, basic appearance rate, avg. and SD of sentiment polarity scores, and the words with the max/min polarity values in CSPD.

カテゴリ X	語彙数	U_{Xave} $\times 10^5$	S_{Xave}	σ_X	最大/最小 E_{XA} の単語 A	評価値 Y での出現率 $U_{XYA} \times 10^5$					感情極性	
						$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$	S_{XA}	E_{XA}
サービス	3,849	74	2.9359	0.2054	親切	1,283	1,923	2,265	5,098	9,401	3.95429	4.95696
					二度と	1,710	365	84	14	12	1.53484	-6.81945
立地	3,173	78	2.9662	0.1710	抜群	50	53	56	163	585	3.90786	5.50419
					悪い	9,068	3,547	2,106	1,134	577	1.84725	-6.54115
部屋	3,866	80	2.9561	0.1817	広々	153	217	322	907	1,940	4.08228	6.19599
					汚い	5,847	1,815	561	223	99	1.53723	-7.80674
設備	3,574	79	2.9688	0.1638	豊富	69	155	212	420	1,023	3.95444	6.01701
					汚い	3,434	1,207	638	233	130	1.74419	-7.47544
風呂	3,013	81	2.9740	0.1873	素晴らしい	182	173	278	649	2,246	4.16928	6.37931
					汚い	3,960	1,289	504	200	73	1.62283	-7.21080
食事	3,242	79	2.9487	0.1976	豊富	788	1,159	1,699	4,569	7,715	4.05749	5.60982
					ひどい	2,057	296	103	49	20	1.52046	-7.22598

$$U_{Xave} = \frac{\sum_{k=1}^N U_{Xk}}{N}.$$

ここで、 N は \mathbf{R}_X 内の感情極性語の語彙数である。この U_{Xave} を基礎出現率と呼ぶ。

- (3) X での感情極性語 A の感情極性スコア S_{XA} を、次の式で計算する。

$$S_{XA} = \frac{\sum_{Y=1}^5 Y(U_{XYA} + U_{Xave})}{\sum_{Y=1}^5 (U_{XYA} + U_{Xave})}.$$

- (4) \mathbf{R}_X 内のすべての感情極性語の X における感情極性スコアを計算した後、次の式で E_{XA} を計算する。

$$E_{XA} = \frac{S_{XA} - S_{Xave}}{\sigma_X}.$$

ここで、 S_{Xave} は、 X におけるすべての感情極性語の感情極性スコアの平均であり、 σ_X は、 X におけるすべての感情極性語の感情極性スコアの標準偏差である。

後述の旧データ (4.1 節) を \mathbf{R} として用いた場合における、各カテゴリ X の CSPD の語彙数、基礎出現率 U_{Xave} 、感情極性スコアの平均 S_{Xave} と標準偏差 σ_X 、感情極性値が最大または最小になった単語 A の各評価値 Y での出現率 U_{XYA} と感情極性スコア S_{XA} および感情極性値 E_{XA} を表 5 に示す。この基礎出現率から、多くの単語の出現率は 0 に近いことが分かる。

上記の計算方法の (3) においては、基礎出現率を足すことにより、 \mathbf{R}_X においてどの評価値においても出現回数が少ないもの、つまり出現率 U_{XA} が基礎出現率を下回るものの感情極性スコアを平均値 $15U_{Xave}/5U_{Xave} = 3$ に近づけることができる。たとえば、ある単語 B が $(U_{X1B}, U_{X2B}, U_{X3B}, U_{X4B}, U_{X5B}) = (0, 0, 0, 0, 0.00001)$ であった場合、この基礎出現率を加えずに感情極性スコアを

表 6 サブカテゴリ辞書の内訳
Table 6 Sub-category breakdown.

サービス	立地	部屋	設備	風呂	食事
言葉遣い・応対 (40)	コンビニ・スーパーへのアクセシビリティ (2)	眺望 (4)	ネット・回線 (15)	温泉 (6)	品数 (6)
食事時の応対 (11)	周辺の飲食店 (6)	開放感 (7)	備品 (28)	露天風呂 (3)	ボリューム (6)
チェックイン時・チェックアウト時の対応 (17)	観光スポットへのアクセシビリティ (6)	清潔感 (25)	館内の設備 (60)	客室風呂 (6)	味 (11)
送迎 (7)	駅 (5)	静けさ (9)	老朽化 (2)	眺望 (4)	朝食 (4)
コスパ (20)	バス (5)	快適さ (12)		設備・備品 (29)	夕食 (2)
おもてなし (9)	駐車 (7)	高級感 (12)			バイキング (3)
	アクセシビリティ (25)	プラン (26)			ジャンル (4)
	周辺の情報 (14)	部屋の設備 (56)			ドリンク (8)

考えた場合には、 $5U_{X5B}/U_{X5B} = 5$ と大きくなってしまふ。しかし提案手法では、そのカテゴリ X の基礎出現率 $U_{X_{ave}}$ が 0.00081 であるとする、 $(5U_{X5B} + 15U_{X_{ave}})/(U_{X5B} + 5U_{X_{ave}}) = (0.00005 + 0.01215)/(0.00001 + 0.00405) = 3.004926$ となることが分かる。これにより、基礎出現率を大きく下回るが、ポジティブな評価値でわずかに使われたものの感情極性スコアが極端に大きくなることを防いでいる。一方、十分に出現率が高い単語についても、多少平均の方へ引っ張られるが、基礎出現率よりも出現率が十分に高ければ、感情極性スコアは平均値より高い値を保つことができる。具体的な例として、表 5 の「風呂」カテゴリにおける「素晴らしい」という単語は、基礎出現率を加えなかった場合の感情極性スコアが 4.3047、加えた場合の感情極性スコアが 4.16928 となった。

さらに、計算方法の (4) により、平均値に近い感情極性スコアを持つ単語の感情極性値を 0 に近くなるようにし、感情極性値のカテゴリごとの分散を 1 にする。表 5 より、感情極性スコアの標準偏差は約 0.2 であったことに注意されたい。これにより、単語間の感情極性値が近くならないようにしている。

3.4 感情特徴量の抽出

CSPD を使用して各カテゴリレビューから特徴量を抽出する。カテゴリレビュー r_X に含まれるすべての感情極性語に、感情極性値を割り当てる。CSPD において感情極性語 A のカテゴリ X での感情極性値が E_{X_A} であるとする、カテゴリレビュー r_X 中で A の次の単語が否定語「ない」の場合、 r_X における A の感情極性値として $E_{X_A} \times (-1)$ を使用する。また、 A が X で「極性値なし」の場合は、 A を X の感情極性語として扱わない。そして、 r_X のすべての感情極性語の感情極性値の平均を計算し、これを r_X の感情特徴量：Sentiment Polarity Feature (SPF) と呼ぶ。SPF の抽出の例を図 2 に示す。

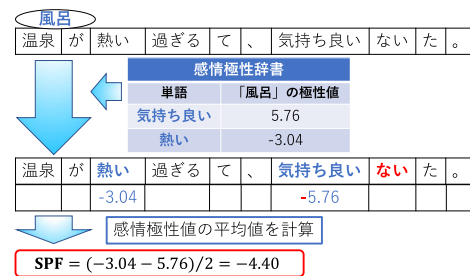


図 2 SPF 抽出の例

Fig. 2 Image of the SPF extraction.

3.5 サブカテゴリへの適用

次に、カテゴリをより細分化したサブカテゴリを考える。本稿では、安部ら [25] によるサブカテゴリと、宿らん [26] のホテル評価基準を基に、人手でサブカテゴリを表 6 のように定義した。そして前述の提案手法を、以下のようにカテゴリ X のサブカテゴリ X_S に適用する。

- (1) カテゴリ辞書で X に含まれる単語のうち、 X_S について述べていると判断しやすい単語を X_S に人手で分類しサブカテゴリ辞書を作成する。表 6 の () 内の数字はそのサブカテゴリに対する登録単語数である。
- (2) カテゴリレビュー r_X を分割し、3.2 節でカテゴリ辞書を用いたのと同様の方法でサブカテゴリ辞書を用いて、サブカテゴリレビュー r_{X_S} を作成する。ただし、サブカテゴリラベル X_S とともに、その上位カテゴリのカテゴリラベル X が付いているもののみを X_S のサブカテゴリレビューとする。以下では、 r_X を r_{X_S} の上位カテゴリレビューと呼ぶ。
- (3) 各単語の感情極性値には、3.3 節でカテゴリを基に作成した CSPD の感情極性値をそのまま用いる。
- (4) 3.4 節と同様の手法で、サブカテゴリレビュー r_{X_S} から SPF を抽出する。

4. 実験評価

4.1 使用データセット

本実験では、楽天トラベルデータセットからの 2 種類の

表 7 各データセットのカテゴリレビューの数
Table 7 Number of category reviews in each data set.

	評価値	1	2	3	4	5	合計	Sub (%)
旧 データ	サービス	21,164	31,956	162,680	292,848	317,222	825,870	91.93
	立地	2,281	16,718	61,225	217,741	365,760	663,725	83.99
	部屋	21,741	51,687	135,273	29,348	342,933	845,117	80.37
	設備	16,892	44,410	140,493	193,902	157,812	553,509	82.78
	風呂	18,169	36,396	56,030	105,716	152,589	368,900	76.27
	食事	13,702	31,076	71,131	167,821	209,811	493,541	89.04
新 データ	サービス	22,579	30,544	105,579	207,398	330,932	697,032	92.18
	立地	2,406	15,282	60,145	165,426	244,721	487,980	86.80
	部屋	23,552	47,038	111,240	233,472	288,757	704,059	78.21
	設備	17,604	40,753	124,569	162,468	141,545	486,939	85.87
	風呂	17,974	35,023	63,905	158,623	248,493	524,018	77.29
	食事	25,278	44,804	96,177	248,635	422,296	837,190	84.36

データセットを用いた。

1 つ目は、2004~2015 年の 1,442,695 件のレビューを含む旧データである。これを文に分割すると、その数は 6,667,090 文となり、そのうちの 38.89% (2,592,830 文) がどのカテゴリにも割り振られなかった。このデータセットのレビューは、3.1 節~3.3 節で説明した方法で CSPD を構築するために使用した。また、この一部は、評価値予測のモデルを作成するためのトレーニングデータとしても用いた。

もう 1 つは、2016~2019 年の 1,914,418 件のレビューを含む新データである。これを文に分割すると、その数は 6,022,256 文となり、そのうちの 35.11% (2,114,490 文) がどのカテゴリにも割り振られなかった。新データの一部は、作成したモデルを評価するためのテストデータに用いた。

新旧データともにおよそ 4 割の文がカテゴリレビューに含まれていない。よって、実際の応用に向けた精度向上のためにはカテゴリ辞書の拡張が必要ではあるが、本稿では抽出されたカテゴリレビューを基に感情極性辞書が作成できるかに焦点を当てているため、カテゴリ辞書の拡張は今後の課題とする。

各データセットの各カテゴリにおける各評価値のカテゴリレビューの数を、表 7 に示す。また、表中の Sub は、各カテゴリのカテゴリレビューのうち、サブカテゴリレビューが抽出されたものの割合を示している。カテゴリレビューに使われた文のうち、どのサブカテゴリレビューにも使われなかった文の数は、旧データで 497,687 文 (12%)、新データで 508,940 文 (13%) であった。カテゴリ辞書によって抽出されたカテゴリレビューに対してのみではあるが、サブカテゴリでその内容を高い割合で網羅できているといえる。

4.2 実験 1: カテゴリレビューの妥当性

まず、被験者へのアンケートによってカテゴリレビュー

表 8 実験 1, 2 の被験者の内訳

Table 8 Breakdown of subjects for Expt. 1-2.

年齢	20 代	30 代	40 代	50 代	60 代以上	合計
男性	1	1	2	2	0	6
女性	1	0	2	2	1	6
合計	2	1	4	4	1	12

の妥当性を評価した。各カテゴリレビュー r_X が、そのカテゴリ X を説明しているか否かを評価する。被験者の内訳を表 8 に示す。旧データから各カテゴリのカテゴリレビューをランダムに 20 件ずつ抽出し、各カテゴリレビューに対して 3 人の被験者を割り当てた。被験者は、与えられたカテゴリレビューを読み、各カテゴリレビューに記載されている内容のカテゴリを選択する。このとき、被験者には、レビューアによる各評価値は与えない。また、被験者は各カテゴリレビューに対して、複数のカテゴリを選択しても良いものとし、また、1 つもカテゴリを選択しないことも可能であるとした。被験者に対する説明は以下のとおりである。

次のホテルレビューの一部を読み、そのレビューが記述していると思うカテゴリ (「サービス」、「立地」、「部屋」、「設備・アメニティ」、「風呂」、「食事」) を選択してください。カテゴリは複数選択可能です。また、1 つも選ばないという選択も可能です。

カテゴリ X のカテゴリレビュー r_X に対して、被験者が選択したカテゴリに X が含まれていた場合に、その被験者は正解したものとする。 r_X に対して、3 人の被験者のうちの少なくとも 1 人が正解した場合に、 r_X は X について妥当であるとした。なお、カテゴリレビューはレビューからカテゴリごとに文を抜粋した文集合であるので、あるレビュー r のカテゴリ X_1 のカテゴリレビュー r_{X_1} が、レビュー r のカテゴリ X_2 のカテゴリレビュー r_{X_2} と文を共

表 9 実験 1 で妥当であるとされたカテゴリレビュー数

Table 9 Number of appropriate category reviews in Expt. 1.

正解者数	サービス	立地	部屋	設備	風呂	食事	合計
1 人	1	1	2	3	1	3	11
2 人	4	0	4	5	4	1	18
3 人	13	15	11	11	14	16	80
合計	18	16	17	19	19	20	109

有する場合がある。このとき、 r_{X_1} に対して被験者は X_2 のみを選択する場合があるが、ここでは r_{X_1} が X_1 を説明していると見なすのが妥当かどうかだけを評価したいため、正解としないことに注意されたい。

1つのカテゴリレビューに対して、被験者が選択したカテゴリの数の平均は 1.522 となった。また、Fleiss の κ 係数 [27] は 0.7457 となった。これは、被験者の一致度を示す評価尺度であり、1 であれば被験者間の回答の完全な一致を示し、0 であれば回答がばらついていることを示している。つまり、各カテゴリレビューの 3 人の被験者の回答はかなり一致していることが分かる。

妥当とされたカテゴリレビューの内訳を表 9 に示す。その割合は全体の 90.83% となった。結果より、カテゴリレビューの妥当性は十分に高いことが分かる。

4.3 実験 2：レビューから受ける印象と評価値の違い

次に、カテゴリレビューから人の目で評価値を予測した際の、実際にレビュアーによって付加された評価値との差を検証する。実験 1 と同時に、被験者には、各カテゴリレビューについて、自身が選択したカテゴリに対してレビュアーがどのような評価値を付けているかを予測してもらった。被験者に対する説明は以下のとおりである。

選択したカテゴリに対してレビュアーがどの評価値を付けているかを予想し、1～5 の段階で選択してください。評価値は、1 が最も評価が低いものとします。

レビュアーによる評価値と被験者が予測した評価値の差を評価するために、表 9 に示したカテゴリレビューのみを対象とし、二乗平均平方根誤差 (RMSE) を計算する。RMSE は以下の式で計算する。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

y_i は正解の評価値、 \hat{y}_i は予測値、 n は予測回数である。ここでは、レビュアーによる評価値を正解とし、被験者による予測評価値を予測値とし、予測回数は被験者による予測評価値の数とする。レビュアーと被験者による 5 段階評価値の RMSE は、1.012 となった。このときの κ 係数は 0.5818 となった。

この結果について、各カテゴリレビューに対する 3 人の

表 10 実験 2 の混同行列

Table 10 Confusion matrix of Expt. 2.

		被験者による予測値の平均				
		1	2	3	4	5
評価値	1	13	6	0	1	1
	2	5	11	4	2	0
	3	1	13	4	4	0
	4	0	2	0	8	11
	5	0	1	0	3	19

被験者による予測値の平均値を四捨五入した値と、レビュアーによる評価値の混同行列を表 10 に示す。また、ポジティブ (評価値 4 または 5) かニュートラルか (評価値 3) かネガティブ (評価値 1 または 2) かの 3 段階評価に変換して評価すると、その RMSE は 0.6455、 κ 係数は 0.683 となり被験者の回答がかなり一致したことが分かる。これらにより、レビュアーと閲覧者の間で、レビューがポジティブな意見を表しているかネガティブな意見を表しているかについては、感覚的知覚に大きな違いはないことが分かる。よって、レビューに含まれる印象を抽出し、それを元に評価値を予測することができれば、一般的なポジティブ/ネガティブの分類を模倣することができるものと考えられる。しかしながら、ポジティブな (またはネガティブな) 意見の中での評価値 4 と 5 (または 1 と 2) の違いに関しては、被験者間で回答のばらつきが大きくなっていることから、個人差が生じることが多く、さらにニュートラルな意見 (評価値 3) の区別は難しいものと考えられる。

4.4 実験 3：カテゴリ評価値予測

ここでは、レビューのカテゴリ評価値の予測を分類問題として解く。分類器として、scikit-learn のロジスティック回帰 [28] を用いる。分類器の学習のために、旧データの各カテゴリのカテゴリレビューからランダムに 70,000 件を選択し、トレーニングデータとした。このトレーニングデータの選択を 5 回行い、そのそれぞれで分類器のモデルを作成する。テストデータは、新データから各カテゴリに対してランダムに抽出された 200,000 件のカテゴリレビューで構成され、5 つのモデルに対して共通のものを用いる。そして、テストデータの各カテゴリレビューの評価値を予測する。

作成した 5 つの各分類器モデルで予測した結果は、以下の 5 つの指標で評価する。答えの評価値を Y 、分類器による予測値を Y_e とすると、

- 正答率：すべてのカテゴリレビューのうち、 $Y = Y_e$ となった割合。
- 適合率： Y_e となったカテゴリレビューのうち、 $Y = Y_e$ となった割合。
- 再現率： Y のカテゴリレビューのうち、 $Y = Y_e$ となった割合。

表 11 感情極性辞書の感情特徴量のみを使用した評価予測の結果
Table 11 Results of the rating prediction by using only SPF of various SPDs.

		サービス	立地	部屋	設備	風呂	食事	平均
正答率	CSPD	0.4666	0.5055	0.4507	0.3780	0.4932	0.5126	0.4678
	PN Table	0.4727*	0.5014 [†]	0.4292 [†]	0.3432 [†]	0.4722 [†]	0.5091 [†]	0.4546 [†]
	JSD	0.4648 [†]	0.5014 [†]	0.4190 [†]	0.3392 [†]	0.4676 [†]	0.5028 [†]	0.4491 [†]
適合率	CSPD	0.2907	0.2409	0.3402	0.2795	0.2766	0.3021	0.2883
	PN Table	0.1596 [†]	0.1003 [†]	0.1564 [†]	0.2149 [†]	0.1384 [†]	0.1784 [†]	0.1580 [†]
	JSD	0.1573 [†]	0.1003 [†]	0.1453 [†]	0.1338 [†]	0.1366 [†]	0.1839 [†]	0.1429 [†]
再現率	CSPD	0.2513	0.2084	0.2594	0.2496	0.2718	0.2692	0.2516
	PN Table	0.2095 [†]	0.2000 [†]	0.2191 [†]	0.2209 [†]	0.2373 [†]	0.2094 [†]	0.2160 [†]
	JSD	0.2088 [†]	0.2000 [†]	0.2089 [†]	0.2111 [†]	0.2453 [†]	0.2148 [†]	0.2148 [†]
F1 値	CSPD	0.2469	0.1552	0.2509	0.2339	0.2466	0.2708	0.2341
	PN Table	0.1660 [†]	0.1336 [†]	0.1765 [†]	0.1948 [†]	0.1724 [†]	0.1619 [†]	0.1675 [†]
	JSD	0.1649 [†]	0.1336 [†]	0.1510 [†]	0.1425 [†]	0.1750 [†]	0.1641 [†]	0.1552 [†]
RMSE	CSPD	1.0297	1.0572	1.0470	1.0663	1.1616	0.9934	1.0592
	PN Table	1.2571 [†]	1.0923 [†]	1.2512 [†]	1.1298 [†]	1.3629 [†]	1.2351 [†]	1.2214 [†]
	JSD	1.2376 [†]	1.0923 [†]	1.3475 [†]	1.0765 [†]	1.3577 [†]	1.2388 [†]	1.2251 [†]

- F1 値：適合率と再現率の調和平均値。
- RMSE：前述のとおり。

適合率、再現率、F1 値は、評価値ごとに計算し、その平均値を用いる。また以下では、5つの分類器モデルによる各指標の平均値を用いて比較する。

4.4.1 感情特徴量

まず、感情特徴量 (SPF) のみを特徴量として使用し、各カテゴリの評価値を予測する。比較する辞書は、CSPD、日本語評価極性辞書 (JSD) [12]、単語感情極性対応表 (PN Table) [13] である。結果を表 11 に示す。結果について t -検定を行い、表中の \dagger は、有意水準 1% で、* は、有意水準 5% で、それぞれ CSPD と比較した際に統計的に有意な差が示されたことを表す。また、3つの辞書のうち、最も良い性能を示した場合を太字で記載している。結果より、CSPD による SPF は、他の辞書による SPF よりもカテゴリ評価値予測において、優れた特徴量であることが分かった。

4.4.2 カテゴリレビューベクトルと感情特徴量

次に、CSPD を用いて各カテゴリレビューから作成した SPF を、そのカテゴリレビューを BERT を用いてベクトル化したものと組み合わせることを考える。さらに、既存の感情極性辞書を用いて作成した SPF と BERT によるベクトルを組み合わせの場合と、ベクトル化手法として fastText を用いた場合とも比較する。

BERT によるベクトルとしては、日本語の事前トレーニング済みモデル [29] を使用し、新旧データセットのすべてのカテゴリレビューのすべての単語をそれぞれ 1,024 次元でベクトル化する。ベクトル作成時には、カテゴリレビューの各単語は表層形のまま入力する。

fastText によるベクトルとしては、Facebook [30] が提供する fastText コードを用いて、各単語を 100 次元でベクトル化する。fastText のトレーニングには、旧データを用

い、そのうち 10 回以上出現した単語のみを用いる。

取得した単語ベクトルを使用して、各カテゴリレビュー r_X について次の特徴ベクトルを作成する：

- BERT： r_X における単語ベクトルの平均ベクトル (1,024 次元)。
- BERT+CSPD：上記の BERT ベクトルに、追加特徴量として CSPD による SPF を結合した 1,025 次元のベクトル。
- BERT+PN：BERT ベクトルに、追加特徴量として PN Table による SPF を結合した 1,025 次元のベクトル。
- BERT+JSD：BERT ベクトルに、追加特徴量として JSD による SPF を結合した 1,025 次元のベクトル。
- fastText： r_X における名詞、動詞、形容詞、形容動詞、および副詞の平均ベクトル (100 次元)。
- fastText+CSPD：上記の fastText ベクトルに、追加特徴量として CSPD による SPF を結合した 101 次元のベクトル。

得られたベクトルを分類器への入力とし、各カテゴリの評価値を予測した。結果を表 12 に示す。結果について t -検定を行い、表中の \dagger は有意水準 1% で、* は有意水準 5% で、それぞれ BERT+CSPD と比較した際に統計的に有意な差が示されたことを表す。また、6つの手法のうち、良い性能を示した場合を太字で記載している。

結果より、CSPD の SPF と BERT の組合せにより、BERT によるベクトルを単独で使用した場合よりも予測精度が高くなった。さらに、BERT と組み合わせる SPF として、他の感情極性辞書によるものを用いた場合と比較すると、CSPD による SPF を用いた方がより大きな精度の向上が認められた。また、文脈依存でないベクトル化手法である fastText を用いた場合でも、CSPD の SPF と組み合わせる

表 12 カテゴリレビューベクトルと感情特徴量を使用した評価予測の結果
Table 12 Results of the rating prediction by using word vectorization and SPF.

		サービス	立地	部屋	設備	風呂	食事	平均
正答率	BERT	0.5486	0.5368*	0.5032 [†]	0.4309	0.5414*	0.5806 [†]	0.5236 [†]
	BERT+CSPD	0.5493	0.5376	0.5058	0.4318	0.5431	0.5826	0.5250
	BERT+PN	0.5484*	0.5372	0.5036 [†]	0.4308	0.5410*	0.5810*	0.5237 [†]
	BERT+JSD	0.5489	0.5369	0.5036 [†]	0.4307	0.5419	0.5807	0.5238 [†]
	fastText	0.5350 [†]	0.5319 [†]	0.4888 [†]	0.4204 [†]	0.5307 [†]	0.5624 [†]	0.5115 [†]
	fastText+CSPD	0.5370 [†]	0.5333 [†]	0.4918 [†]	0.4221 [†]	0.5324 [†]	0.5656 [†]	0.5137 [†]
適合率	BERT	0.4488*	0.4098	0.4305 [†]	0.4020	0.4192 [†]	0.4546*	0.4275*
	BERT+CSPD	0.4508	0.4162	0.4352	0.4032	0.4234	0.4581	0.4311
	BERT+PN	0.4506	0.4356	0.4331 [†]	0.4019	0.4189 [†]	0.4552	0.4325
	BERT+JSD	0.4490	0.4443*	0.4330	0.4021	0.4203*	0.4540	0.4338
	fastText	0.4216 [†]	0.4171	0.4006 [†]	0.3740 [†]	0.4132	0.4389 [†]	0.4109 [†]
	fastText+CSPD	0.4253 [†]	0.3502 [†]	0.4048 [†]	0.3798 [†]	0.4134 [†]	0.4434 [†]	0.4028 [†]
再現率	BERT	0.3976 [†]	0.2655 [†]	0.3744*	0.3352	0.3874	0.3964	0.3594 [†]
	BERT+CSPD	0.4023	0.2710	0.3799	0.3360	0.3887	0.4021	0.3633
	BERT+PN	0.3997	0.2666*	0.3734*	0.3345	0.3879*	0.3989	0.3602 [†]
	BERT+JSD	0.4002	0.2643 [†]	0.3735*	0.3362	0.3897	0.3997	0.3606 [†]
	fastText	0.3322 [†]	0.2439 [†]	0.3094 [†]	0.2898 [†]	0.3393 [†]	0.3295 [†]	0.3074 [†]
	fastText+CSPD	0.3404 [†]	0.2464 [†]	0.3156 [†]	0.2929 [†]	0.3445 [†]	0.3418 [†]	0.3136 [†]
F1 値	BERT	0.4115*	0.2693 [†]	0.3895 [†]	0.3477	0.3923 [†]	0.4158*	0.3710 [†]
	BERT+CSPD	0.4153	0.2779	0.3952	0.3488	0.3952	0.4213	0.3756
	BERT+PN	0.4146	0.2710*	0.3898*	0.3473	0.3926 [†]	0.4183	0.3723 [†]
	BERT+JSD	0.4142	0.2674 [†]	0.3899 [†]	0.3490	0.3948	0.4182	0.3723 [†]
	fastText	0.3476 [†]	0.2332 [†]	0.3189 [†]	0.2850 [†]	0.3532 [†]	0.3505 [†]	0.3148 [†]
	fastText+CSPD	0.3539 [†]	0.2372 [†]	0.3265 [†]	0.2897 [†]	0.3579 [†]	0.3645 [†]	0.3216 [†]
RMSE	BERT	0.9362 [†]	0.9059 [†]	0.9766 [†]	1.0309 [†]	0.9976 [†]	0.9039 [†]	0.9585 [†]
	BERT+CSPD	0.9270	0.8999	0.9635	1.0236	0.9878	0.8887	0.9484
	BERT+PN	0.9312*	0.9038*	0.9755 [†]	1.0297*	0.9956 [†]	0.8981 [†]	0.9557 [†]
	BERT+JSD	0.9312 [†]	0.9061 [†]	0.9755 [†]	1.0290*	0.9918	0.8971 [†]	0.9551 [†]
	fastText	1.0214 [†]	0.9340 [†]	1.0278 [†]	1.0532 [†]	1.0308*	0.9802 [†]	1.0079 [†]
	fastText+CSPD	1.0016 [†]	0.9274 [†]	1.0059 [†]	1.0442 [†]	1.0207 [†]	0.9543 [†]	0.9924 [†]

表 13 予測値が大きく外れたカテゴリレビューの内訳
Table 13 Breakdown of the category reviews with significantly incorrect prediction.

評価値	予測値	サービス	立地	部屋	設備	風呂	食事	合計
1	5	744	355	500	311	831	703	3,444
1	4	817	404	1,320	1,182	1,241	977	5,941
2	5	1,381	1,907	1,361	1,077	2,111	1,340	9,177
4	1	395	4	415	236	629	394	2,073
5	2	104	69	193	171	534	211	1,282
5	1	162	2	125	68	255	161	773
合計		3,603	2,741	3,914	3,045	5,601	3,786	22,690

ことにより、精度が向上している。fastText よりも BERT を用いることで予測精度が向上していることも分かる。

4.5 実験 4：予測値と評価値の差が 3 以上のものについて
レビューアーによる評価値と BERT+CSPD による予測値の差が大きくなったカテゴリレビューについて観察する。表 13 に評価値と予測値の差が 3 以上となったもののカテゴリごとの数を示す。これにより、どのカテゴリでも、評

価値がネガティブなものをポジティブに予測したケースが多いことが分かる。

この中から、各カテゴリについて、評価値が 1 で予測値が 5 となったカテゴリレビューからランダムに 10 件と、評価値が 5 で予測値が 1 となったカテゴリレビューからランダムに 10 件の計 20 件を抽出する。ただし、「立地」のカテゴリに関しては評価値が 5 で予測値が 1 となったものの数が十分ではなかったため、以下の内訳で補っている。

表 14 実験 4 の被験者の内訳

Table 14 Breakdown of subjects for Expt. 4.

年齢	20 代	30 代	40 代	50 代	60 代以上	合計
男性	5	1	1	0	0	7
女性	0	0	1	1	1	3
合計	5	1	2	1	1	10

表 15 BERT+CSPD が正しいとされた場合の内訳

Table 15 Breakdown of cases that BERT+CSPD is correct.

評価値	サービス	立地	部屋	設備	風呂	食事	合計
ポジティブ	7	10	8	8	7	6	46
ネガティブ	1	2	1	4	1	2	11
合計	8	12	9	12	8	8	57

- 評価値 5 で予測値 2 となったもの 4 件
- 評価値が 4 で予測値が 1 となったもの 4 件

これらの各カテゴリレビューに、表 14 に示す被験者から 3 人を割り当て、評価値と予測値のどちらが正しいかを判定してもらう。ただし、被験者には具体的な評価値は与えず、各カテゴリレビューがポジティブな内容であるかネガティブな内容であるかを選択してもらう。被験者に対する説明は以下のとおりである。

次のホテルに対するレビューを読み、指定されたカテゴリに対して、ポジティブまたはネガティブのどちらよりの意見が書かれているかを選択してください。

BERT+CSPD による予測値の方が正しいと判定されたカテゴリレビューの内訳を表 15 に示す。この回答における κ 係数は 0.6304 となり、3 人の被験者による回答はかなり一致していた。BERT+CSPD による予測値の方が正しいと判定されたのは全体の 47.5% となった。また、評価値がネガティブなものに関してはその評価値のままの方が被験者に対して正しい印象を与えているが、評価値がポジティブなものに関しては、被験者に対してネガティブな印象を与えるものが多く (76.7%)、それらに関しては BERT+CSPD を用いることで、ある程度、検出ができていたといえる。よって、予測値によってレビューアの評価値を置き換えられるほどの精度には達していないが、1 章で述べたようにレビューアに確認を促すには十分な役割を果たせるのではないかと考える。

4.6 実験 5: サブカテゴリ

ここでは、「食事」のカテゴリを細分化したサブカテゴリを用いて、それらに対する 5 段階および 3 段階の評価値予測の評価実験を行った。

まず、新データからテストデータを作成する。サブカテゴリについては、レビューアによる評価値が記載されていないため、被験者へのアンケートによって、サブカテゴリの評価値の答えを作成した。被験者の内訳を表 16 に示す。

表 16 実験 5 の被験者の内訳

Table 16 Breakdown of subjects for Expt. 5.

年齢	20 代	30 代	40 代	50 代	60 代以上	合計
男性	0	1	0	2	0	3
女性	1	3	0	1	1	6
合計	1	4	0	3	1	9

新データから「食事」カテゴリのサブカテゴリに属する 480 件のサブカテゴリレビューを抽出し、各サブカテゴリレビューに対して 3 人の被験者を割り当てた。被験者は、与えられたサブカテゴリレビューを読み、各サブカテゴリレビューに記載されている内容のサブカテゴリを選択し、それらのサブカテゴリに対してレビューアが付加するであろう評価値を 5 段階で予測した。このとき、被験者は各サブカテゴリレビューに対して、複数のサブカテゴリを選択しても、1 つも選択しなくても良いものとした。各サブカテゴリレビューについて、少なくとも 1 人の被験者が選択したサブカテゴリには、予測値が付くことになる。複数人の被験者たちが 1 つの同じサブカテゴリを選択した場合には、そのサブカテゴリを選んだ被験者達による予測値の平均を四捨五入した値をそのサブカテゴリに対する予測値とする。被験者への説明は以下のとおりである。

以下のレビューの一部はカテゴリ「食事」に分類されます。カテゴリ「食事」は、「品数」、「ボリューム」、「味」、「朝食」、「夕食」、「バイキング」、「食のジャンル」、「ドリンク」の 8 つのサブカテゴリに分かれます。レビューの一部を読み、そのレビューが記述していると思うサブカテゴリを選択してください。また、そのサブカテゴリに対してレビューアがどの評価値を付けているかを予想し、1~5 の段階で選択してください。サブカテゴリは複数選択可能です。1 つも選ばないという選択も可能です。評価値は 1 が最も評価が低いものとしします。

ここでは 3 段階評価の評価値予測も行うので、被験者が予測した評価値 1 と 2 をまとめてネガティブ、評価値 3 をニュートラル、評価値 4 と 5 をポジティブな評価と見なし 3 段階に変換し、そのサブカテゴリレビューの評価値の答えとした。

4.6.1 サブカテゴリレビューの妥当性

これを用いて、まずサブカテゴリ辞書によるサブカテゴリ分類の妥当性を評価した。実験 1 と同様の方法で、あるサブカテゴリレビュー r_{X_S} のサブカテゴリ X_S が、被験者が選択したサブカテゴリに含まれていた場合に、その被験者は正解したものとし、3 人の被験者のうちの少なくとも 1 人が正解した場合に、そのサブカテゴリレビュー r_{X_S} は妥当であるとした。結果として、妥当とされたサブカテゴリレビュー数は 446 件で、これは全体の 92.92% である。そ

のうちの75件が2人の被験者が正解し、341件が3人の被験者が正解した。また、このときの κ 係数は0.7746となった。つまり、3人の被験者の回答はかなり一致していることが分かる。よって、サブカテゴリレビューの妥当性は十分に高いことが分かる。

4.6.2 作成されたテストデータ

妥当とされた446件のサブカテゴリレビューについて、被験者による予測値ごとのサブカテゴリレビューの数を表17に示す。この評価値の回答における κ 係数は0.5693となり、被験者の回答は適度に一致している。また、3段階に変換したときの回答における κ 係数は0.726となり、被験者間で判断がかなり一致しているといえる。つまり、実験2と同様に、ポジティブであるかネガティブであるかの判断はあまり被験者によって違いがないが、ポジティブな中での評価値4と5、ネガティブな中での評価値1と2の区別には個人差があることが分かる。

4.6.3 サブカテゴリ評価値予測

続いて、サブカテゴリについての5段階および3段階の評価値予測の評価実験を行った。旧データの各サブカテゴリのサブカテゴリレビューから、ランダムに70,000件を選択し、トレーニングデータとした。トレーニングデータの評価値の答えとしては、サブカテゴリレビュー r_{X_S} のサブカテゴリ X_S の属するカテゴリ X （ここでは「食事」）に対してレビューアが付加した評価値を用いた。このことの影響についての分析は今後の課題であるが、カテゴリというごく限られた内容についてのことであるので影響は小さいものとする。このトレーニングデータの選択を5回行い、そのそれぞれで分類器のモデルを作成し、テストデータの各サブカテゴリレビューについて評価値を予測する。実験3と同様に、5つの分類器モデルによる各精度指標の平均値を用いて比較を行う。

本実験では、4.4.2項と同様に、CSPDによるSPFとBERTによるベクトルを用いた予測精度の比較を行った。BERTの日本語の事前トレーニング済みモデル[29]を使用し、新旧データセットのすべてのカテゴリレビューのすべての単語をそれぞれ1,024次元でベクトル化する。そして、取得した単語ベクトルを使用して、各サブカテゴリレビュー r_{X_S} について次の3つの特徴ベクトルを作成する：

- BERT： r_{X_S} における単語ベクトルの平均ベクトル（1,024次元）。
- BERT+CSPD：上記の平均ベクトルに、追加特徴量として r_{X_S} のSPFを結合した1,025次元のベクトル。
- BERT+CSPD+CSPF：上記のBERT+CSPDのベクトルに、追加特徴量として r_{X_S} の上位カテゴリレビュー r_X のSPF（CSPFと呼ぶ）を結合した1,026次元のベクトル。

サブカテゴリに対する評価値は、サブカテゴリに対する記述だけではなく、その上位カテゴリに対する記述にも依存

表17 テストデータのサブカテゴリレビューの数

Table 17 Number of subcategory reviews in the test data.

評価値	1	2	3	4	5	合計
品数	5	12	3	15	16	51
ボリューム	0	2	5	18	31	56
味	1	0	4	27	28	60
朝食	0	7	9	24	18	58
夕食	6	14	5	16	16	57
バイキング	4	9	6	20	20	59
ジャンル	3	9	4	18	14	48
ドリンク	6	6	6	24	15	57

すると考えられる。よって、3つ目のベクトルでは、CSPFを加えた。

結果を表18に示す。この結果について、8つのサブカテゴリのすべての試行に対する t -検定を行い、表中の「平均」の欄に結果を示す。 \dagger は有意水準1%で、 $*$ は有意水準5%で、それぞれBERT+CSPD+CSPFを用いた場合と比較した際に統計的に有意な差が示されたことを表す。

まず、5段階評価の結果については、平均的にはBERT単体の方がほぼすべての指標で良い結果となっていた。しかし、「夕食」ではBERT+CSPD+CSPFがすべての指標で、「味」と「ドリンク」ではBERT+CSPDが3つの指標で良い結果となるなど、サブカテゴリによってばらつきがあった。

次に、3段階評価の結果について考察する。「味」と「ボリューム」ではBERT単体の方が良い結果となったが、他のカテゴリではBERT+CSPD+CSPFがおおむね良い結果となっている。ここで表17を見てみると、これらの2つサブカテゴリでは、ネガティブが正解となるものが非常に少ない。つまり、BERT単体では、評価値4と5の区別は非常によくできているが、それらと評価値1～3のポジティブでないものを区別することができていないことが分かる。一方、BERT+CSPD+CSPFでは、評価値4と5の区別はあまりうまくいっていないが、それらとネガティブなものとの区別はよくできていることが分かる。また、サブカテゴリについては、データとしてレビューアの評価値がなく、被験者による評価値予測を正解としている。しかし4.6.2項で述べたように、被験者による5段階の評価値にはある程度の個人差が表れている。よって、評価値の1程度の差を区別できることよりも、ポジティブであるかネガティブであるかの差を区別できることの方が、1章で示したレビューアへの注意喚起といった応用のためには有意義であると考えられる。

5. おわりに

本稿では、ホテルのレビューから各カテゴリに対する評価値を予測する方法を提案した。各カテゴリにおける単語の出現率を使用して、各カテゴリに特化した感情極性辞書

表 18 サブカテゴリの評価値予測の結果

Table 18 Results of the rating prediction for sub-categories.

			品数	ボリューム	味	朝食	夕食	バイキング	食のジャンル	ドリンク	平均
5	正答率	BERT	0.5020	0.5750	0.5900	0.6172	0.3263	0.4644	0.4333	0.3965	0.4881[†]
		BERT+CSPD	0.4941	0.5643	0.5967	0.5621	0.3404	0.4407	0.4250	0.4140	0.4797
		BERT+CSPD+CSPF	0.4745	0.5286	0.5767	0.5897	0.3614	0.4576	0.4250	0.3614	0.4719
	適合率	BERT	0.5148	0.4787	0.4279	0.6194	0.2325	0.5191	0.5639	0.3095	0.4582
		BERT+CSPD	0.5069	0.4271	0.4095	0.4867	0.3067	0.4387	0.5377	0.3425	0.4320
		BERT+CSPD+CSPF	0.4321	0.4302	0.4244	0.5116	0.3203	0.4398	0.5074	0.4120	0.4347
	再現率	BERT	0.4153	0.3960	0.3728	0.4742	0.2381	0.3618	0.4130	0.3023	0.3717[†]
		BERT+CSPD	0.4200	0.3475	0.3764	0.3799	0.2564	0.3342	0.3894	0.2997	0.3504
		BERT+CSPD+CSPF	0.3787	0.3314	0.3655	0.3887	0.2763	0.3460	0.3552	0.2787	0.3400
	F1 値	BERT	0.4132	0.4145	0.3688	0.4857	0.1842	0.3628	0.4098	0.2871	0.3658*
		BERT+CSPD	0.4125	0.3569	0.3662	0.3888	0.2076	0.3255	0.3788	0.2886	0.3406
		BERT+CSPD+CSPF	0.3742	0.3430	0.3607	0.3991	0.2385	0.3439	0.3492	0.2850	0.3367
3	正答率	BERT	1.1426	0.7254	0.7605	1.0227	1.5411	1.2489	1.0848	1.5387	1.1331
		BERT+CSPD	1.1508	0.7989	0.7504	1.1122	1.5853	1.2536	1.1354	1.4898	1.1596
		BERT+CSPD+CSPF	1.1037	0.8252	0.7826	1.0739	1.4819	1.2648	1.1261	1.4992	1.1447
	適合率	BERT	0.7137	0.9036	0.9267	0.7621	0.6526	0.7627	0.7583	0.6702	0.7704*
		BERT+CSPD	0.6980	0.9036	0.9200	0.7586	0.6561	0.7559	0.7792	0.6737	0.7695 [†]
		BERT+CSPD+CSPF	0.7216	0.9000	0.9167	0.7655	0.6702	0.7763	0.7958	0.6667	0.7776
	再現率	BERT	0.5341	0.7037	0.4139	0.7693	0.5442	0.7961	0.5898	0.4066	0.5947
		BERT+CSPD	0.5241	0.7670	0.4138	0.7241	0.5485	0.7676	0.5744	0.4104	0.5913
		BERT+CSPD+CSPF	0.5426	0.7689	0.3107	0.7527	0.6341	0.7999	0.5987	0.4952	0.6128
	F1 値	BERT	0.4481	0.5400	0.3833	0.4593	0.4200	0.5154	0.4556	0.3735	0.4494
		BERT+CSPD	0.4324	0.5600	0.3809	0.4418	0.4233	0.5180	0.4833	0.3697	0.4512
		BERT+CSPD+CSPF	0.4577	0.5306	0.3333	0.4608	0.4767	0.5359	0.5056	0.3970	0.4622
3	正答率	BERT	0.4537	0.5781	0.3900	0.4788	0.3954	0.5525	0.4683	0.3481	0.4581
		BERT+CSPD	0.4342	0.6048	0.3888	0.4575	0.4027	0.5548	0.4951	0.3441	0.4603
		BERT+CSPD+CSPF	0.4678	0.5840	0.3216	0.4818	0.4714	0.5813	0.5213	0.3972	0.4783
	適合率	BERT	0.8895	0.3377	0.2704	0.7332	1.0309	0.8268	0.7683	1.0070	0.7283 [†]
		BERT+CSPD	0.9244	0.3696	0.3132	0.7351	1.0027	0.8488	0.7380	0.9998	0.7375 [†]
		BERT+CSPD+CSPF	0.8577	0.4021	0.2887	0.7154	0.9856	0.7940	0.6825	0.9602	0.7076

を作成した。提案した感情極性辞書は、BERT によるベクトルと組み合わせることで、BERT のみを使用する場合よりも優れた結果を示した。

今後の課題として、評価値予測精度の向上のために、カテゴリ辞書、サブカテゴリ辞書の拡充を行い、カテゴリレビュー、サブカテゴリレビューの抽出の再現率を高める必要がある。また、CSPD の感情極性値を加える文ベクトルを作成するのに、BERT による平均ベクトルよりも精度が良いといわれている Sentence-BERT [31] を用いることも考えられる。さらに、CSPD の適用方法として、文ベクトルまたは単語ベクトルの学習時にその感情極性値を考慮した転移学習も検討する予定である。

謝辞 本研究では、国立情報学研究所の 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」を利用した。

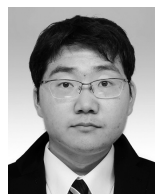
参考文献

- [1] Shibata, A., Kamei, S. and Nakano, K.: Category-oriented Sentiment Polarity Dictionary for Rating Prediction of Japanese Hotels, *Proc. 11th International Workshop on Advances in Networking and Computing* (2020).
- [2] 楽天グループ株式会社：楽天トラベル，入手先 (<https://travel.rakuten.com/>).
- [3] TripAdvisor: available from (<https://www.tripadvisor.jp/>).
- [4] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR*, Vol.abs/1810.04805 (2018).
- [5] 外山洋太，三輪 誠，佐々木裕：文書・文間及びカテゴリ間の関係を考慮したレーティング予測，言語処理学会第 22 回年次大会発表論文集，pp.158-161 (2016).
- [6] 張 博，白井清昭：レビューテキストの書き手の評価視点に対する評価点の推定，言語処理学会第 23 回年次大会発表論文集，pp.803-806 (2017).
- [7] Lamdabhin, M., Bouzaidi, L. and Zahir, J.: A Model for Aspect Based Sentiment Analysis in Hotel Reviews using RNNs, *Proc. 1st UM6P Spring School on Complexity Analysis of Industrial Systems and Advanced Modeling*

- (2019).
- [8] Le, Q. and Mikolov, T.: Distributed representations of sentences and documents, *Proc. 31th International Conference on Machine Learning*, pp.II-1188–II-1196 (2014).
 - [9] Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality, *Proc. 26th International Conference on Neural Information Processing Systems*, pp.3111–3119 (2013).
 - [10] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching word vectors with subword information, *Trans. Association for Computational Linguistics*, Vol.5, pp.135–146 (2017).
 - [11] Thelwall, M., Buckley, K., Paltoglou, G., et al.: Sentiment strength detection in short informal text, *Journal of the American Society for Information Science and Technology*, Vol.61, No.12, pp.2544–2558 (2010).
 - [12] 東北大学 乾・鈴木研究室: 日本語評価極性辞書, 入手先 (<https://www.nlp.ecei.tohoku.ac.jp/research/open-resources/>).
 - [13] Takamura, H., Inui, T. and Okumura, M.: Extracting semantic orientations of words using spin model, *Proc. 43rd Annual Meeting of the Association for Computational Linguistics*, pp.133–140 (2005).
 - [14] 森田晋也, 白井靖人: 分野別単語感情極性辞書の作成及び評価, 第 80 回全国大会講演論文集, Vol.2018, No.1, pp.317–318 (2018).
 - [15] Hu, M. and Liu, B.: Mining and summarizing customer reviews, *Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004).
 - [16] Popescu, A.-M. and Etzioni, O.: Extracting Product Features and Opinions from Reviews, *Proc. Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp.339–346 (2005).
 - [17] Fang, X. and Zhan, J.: Sentiment analysis using product review data, *Journal of Big Data*, pp.1–14 (2015).
 - [18] Lu, Y., Castellanos, M., Dayal, U. and Zhai, C.: Automatic construction of a context-aware sentiment lexicon: an optimization approach, *Proc. 20th International Conference on World Wide Web*, pp.347–356 (2011).
 - [19] Almatrneh, S. and Gamallo, P.: Automatic construction of domain-specific sentiment lexicons for polarity classification, *Proc. 15th International Conference on Practical Applications of Agents and Multi-Agent Systems*, pp.175–182 (2017).
 - [20] 小林のぞみ, 乾健太郎, 松本裕治ほか: 意見抽出のための評価表現の収集, 自然言語処理, Vol.12, No.3, pp.203–222 (2005).
 - [21] 楽天グループ株式会社: 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), DOI: 10.32130/idr.2.0 (2014).
 - [22] Takuma, K., Yamamoto, J., Kamei, S. and Fujita, S.: A Hotel Recommendation System Based on Reviews: What Do You Attach Importance To?, *Proc. 4th International Symposium on Computing and Networking*, pp.710–712 (2016).
 - [23] 楽天グループ株式会社: 筑波大学文単位評価極性タグ付きコーパス. 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), DOI: 10.32130/idr.2.6 (2014).
 - [24] 打田智子: Janome(ja), 入手先 (<https://mocabeta.github.io/janome/>).
 - [25] 安部 克, 酒井 聖, 松波友稀, 中島伸介: レビュー評価項目別スコアを用いたホテル推薦手法の提案, 第 11 回デー

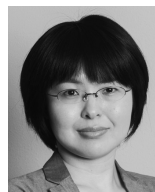
- タ工学と情報マネジメントに関するフォーラム (2019).
- [26] 宿らん調査員: 入手先 (<http://www.yadoran.jp>) (参照 2019-09).
 - [27] Fleiss, J.L.: Measuring nominal scale agreement among many raters, *Psychological Bulletin*, Vol.76, No.5, pp.378–382 (1971).
 - [28] Cox, D.R.: The regression analysis of binary sequences, *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol.20, No.2, pp.215–232 (1958).
 - [29] 京都大学黒橋・河原・村脇研究室: BERT 日本語 Pre-trained モデル, 入手先 (http://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese).
 - [30] Facebook: Github – facebookresearch/fasttext: Library for fast text representation and classification, available from (<https://github.com/facebookresearch/fastText>).
 - [31] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *CoRR*, Vol.abs/1908.10084 (2019).

柴田 諒人



広島大学大学院工学研究科博士課程前期在学中。2019 年広島大学工学部第二類情報工学プログラム卒業。2021 年同大学大学院工学研究科博士課程前期修了。同年ジェイアール東海情報システム株式会社入社。

亀井 清華 (正会員)



広島大学大学院先進理工系科学研究科准教授。2006 年広島大学大学院博士課程後期修了。博士 (工学)。同年鳥取環境大学環境情報学部助手。2008 年広島大学大学院工学研究科情報工学専攻助教。2012 年同准教授。分散アルゴリズムおよび推薦システムの研究に従事。電子情報通信学会, IEEE, ACM 各会員。

中野 浩嗣 (正会員)



広島大学大学院先進理工系科学研究科教授。1992 年大阪大学博士後期課程修了。博士 (工学)。同年日立基礎研究所研究員。1995 年名古屋工業大学講師。2001 年北陸先端科学技術大学院大学助教授。2003 年広島大学工学研究科教授。並列分散処理の研究に従事。電子情報通信学会, IEEE 各会員。

(担当編集委員 佃 洸撰)