DEIM Forum 2021 H11-4

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

小橋　賢介†　　雨宮　佑基†　　酒井　哲也†

† 早稲田大学基幹理工学研究科情報理工・情報通信専攻　〒 169–8555 東京都新宿区大久保 3–4–1
E-mail: †lelouch-ken@ruri.waseda.jp, ††yukiamemiya@fuji.waseda.jp, †††tetsuyasakai@acm.org

あらまし　ショッピングサイトで閲覧できる商品レビューには実際に商品を購入したユーザの体験に基づく情報が含まれているため，購入者が商品選定を行ったり，商品を販売する企業が商品の評判を分析したりする際に非常に有用である．その一方で，このような商品選定や評判分析では目的の商品だけでなく類似商品におけるレビューも参照する必要があり，これはユーザにとって大きな負荷となる．こういった負荷を軽減させるための研究分野の一つとして，商品選定に有用なレビューを自動的に分類することを目的とした Helpful Review Prediction という分類タスクが注目されている．我々はこのタスクにおいて，商品の優劣関係を容易に把握することができる比較文に着目し，比較文を含むレビューを取り出した後で分類を行うことでさらなる分類精度の向上につながるという仮説を立てた．そこで本研究では，あらかじめ比較文を含むレビューをフィルタリング（抽出）した場合とそうでない場合について，Helpful Review Prediction の分類精度を比較する実験を行った．その結果，比較文を含むレビューをフィルタリングすることが有効であるということが示された．

キーワード　Helpful Review Prediction，評判分析，テキストマイニング，情報抽出，E-commerce

1　は じ め に

インターネットの急激な普及に伴い，Amazon[1]や Yahoo! ショッピング [2]，楽天市場 [3]などのショッピングサイトにおいて，ユーザが実際の商品購入者によって投稿されたレビューを閲覧する機会が増加している．消費者庁の 2017 年消費者意識基本調査によると，ユーザは商品を選定する際に価格や機能，安全性など様々な観点を考慮しているという [4]．例えば，購入したい商品がカメラである場合，写真を撮るなどの基本機能や価格はショッピングサイトの商品詳細ページに記載されているため容易に把握できるが，カメラのボタンの配置など詳細な機能や使い心地，苦情や要望への対応を把握することは困難である．一方で，実際の商品購入者によって投稿されたレビューは信頼性が高く，ユーザが欲している情報量が多いため，商品購入者が商品選定をする際にレビューを利用することは非常に有用である．さらに，商品販売している企業にとっても，レビューは自社商品や競合他社の商品の評判を分析するための重要なリソースとなる．

商品購入の意思決定や商品の評判分析には，目的の商品だけでなく類似商品のレビューを閲覧することが必要である．しかし，各商品に対して公開されているレビュー数は膨大であり，商品選定や評判分析のためにレビューを閲覧することは多くの時間を要するため，ユーザや企業にとって大きな負荷となる．このような状況では，ユーザは負荷を処理することができず，ユーザの購買意欲の低下やそれによるショッピングサイト側の販売機会の損失に繋がる可能性がある．そのため，レビューを閲覧する負荷を軽減するために，数多くのレビューの中から商品選定に有用なレビューを抽出し，ユーザに提示するシステムの構築が必要であると考えられる．

ここで，商品選定に有用なレビューを抽出することでユーザの負荷を軽減する 2 種類の研究の例を挙げる．1 つ目は，比較文が含まれるレビュー中から，購入を検討している商品と類似商品との比較関係を抽出する比較文抽出タスクである．2 つ目は，与えられたレビューに対して *Helpful* か否かの 2 値分類を行うことで，疑似的に *Helpful* なレビューを分類する Helpful Review Prediction というタスクである．なお，*Helpful* なレビューとは，そのレビューを閲覧したユーザの多くから役に立ったと評価されたレビューを意味する．このタスクは，レビューが投稿されてからの時間があまり経過していないレビューの投票数が少なくなり，そのレビューがユーザにとって役立つかどうかを判断することが難しいという課題の解決を目的としている．しかし，その分類精度には未だ改善の余地があり，我々は商品選定に有効である比較文抽出タスクを Helpful Review Prediction に組み込むことで，与えられたレビューが *Helpful* であるかどうかの分類精度がより向上すると考えた．そこで本研究では，あらかじめ比較文を含むレビューをフィルタリングを挿入した場合とそうでない場合について，Helpful Review Prediction のタスクに対する精度を比較する実験を行った．この実験により，Helpful Review Prediction を行う前に比較文を含むレビューをフィルタリングすることが有効であるということが示された．

1 : https://www.amazon.com/
2 : https://shopping.yahoo.com/
3 : https://www.rakuten.com/
4 : https://www.caa.go.jp/policies/policy/consumer_research/white_paper/2018/white_paper_121.html

2 関連研究

2.1 比較文抽出タスク

レビューから比較文を抽出することで，商品同士の比較関係から商品の特性を把握することができるため，膨大なレビューの中から商品選定を行うユーザの負荷を減らすことができると考えられる．そのため，レビューから比較文を抽出する研究が数多く行われている．Jindal ら [1] は，Class Sequential Rule (CSR) と機械学習を利用してレビューが比較文であるか否かの 2 値分類を行った．さらに，学習データを作成する際の比較文であるか否かの手動ラベル付けの負荷を軽減するため，疑似的にラベル付けを行う単語レベルや品詞レベルのルールを作成した．

Xu ら [2] は，1 文ごとでしか比較商品や比較ワードを抽出することができなかった Conditional Random Fields (CRF) を複数文に対しても比較商品や比較ワードを抽出できるような two-level CRF を提案した．そして，その two-level CRF 用いてスマートフォンに関するレビューからどちらの商品が優れているかを示す比較方向や比較商品，比較ワード，特徴を以下の例のように抽出した．

レビュー例：
“Compared with Nokia N95, iPhone has a better camera.”
比較関係の抽出例：
>(Nokia N95, iPhone, camera, better)

Danone ら [3] は，Xu らの手法を用いてレストラン口コミサイト [5]のレビューから比較方向や比較商品，比較ワード，特徴を抽出した．さらに，抽出した結果を棒グラフやレーダーチャート，表によって視覚化を行い，どのグラフが見やすいかという観点でユーザーからアンケートを取った．その結果，棒グラフが最も見やすい可視化手法であることが示された．

2.2 Helpful Review Prediction

Helpful Review Prediction とは，与えられたレビューが *Helpful* であるか否かを機械学習を利用して判定するタスクのことであり，Helpful Review Prediction に関する研究は以下のように数多く行われている．

Haque ら [4] は，商品レビューを lexical，structural，semantic の観点で特徴量生成を行うだけでなく，Flesch reading ease score [5] を利用してレビュー全体の読みやすさスコアも特徴量に追加した．これらの特徴量と決定木モデルを用いて *Helpful* であるか否かの 2 値分類を行った．

Mukherjee ら [6] は，教師なし学習モデルである HMM-LDA を用いて Helpful Review Prediction を行った．過去の研究ではドメイン固有の特徴量を利用していたが，Mukherjee らはユーザの知識や書き方，レビューの一貫性などのドメイン非依存の特徴量を利用することで，ドメインによらずに Helpful Review Prediction を行うことを可能にした．

5：https://www.yelp.com/

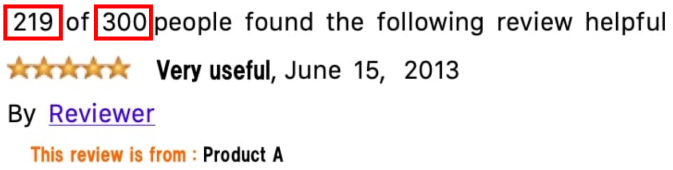

図 1　レビューサンプル

Martin ら [7] は，ホテル予約サイトのレビュー中で表されている感情は *Helpful* なレビューかどうかの良い判断基準となる可能性があると考え，レビュー中で表されている感情を特徴量としてランダムフォレストに学習させ，Helpful Review Prediction を行った．Malik ら [8] も同様に，Amazon のレビュー内で表されている感情を特徴量として用いた．より具体的には，ポジティブな感情（驚き・信頼・期待・喜び）とネガティブな感情（不安・嫌悪・怒り・悲しみ）を特徴量として加え，これらの特徴量を用いてディープニューラルネットワークによる分類を行った．

さらに先行研究では，Helpful Review Prediction を行うためにトレーニングデータセットを作成する過程で，レビューが *Helpful* か否かについてのラベル付けを行った研究も存在する．図 1 に示すように，Aamazon のレビューにはレビューが *Helpful* かどうかを投票できる機能があり，そのレビューには，X 人中 Y 人が役に立った (*Helpful*) と投票したかが記載されている．ここで，*Helpful* なレビューとは，投票した全体の人数に対し *Helpful* と投票した人数の割合が高いレビューを指し，その割合を Helpful Probability と呼ぶ．Ghose ら [9] は，2 人のアノテータにのレビューが *Helpful* であるか否かに関するラベル付けを行った．その結果と Helpful Probability が 0.6 以上であるレビューを *Helpful* と設定した場合，アノテータのラベル付けと Helpful Probability によるラベル付けの誤差が最も少なくなる．その結果を利用して，Helpful Probability が 0.6 以上であるレビューを *Helpful*，0.6 未満であるレビューを *Unhelpful* と設定した．

3 比較文フィルタリング

本研究では，商品選定に有効である比較文抽出タスクを Helpful Review Prediction に組み込むことで，与えられたレビューが *Helpful* であるかの 2 値分類の精度がより向上するという仮説を立てた．そこで，比較文を含むレビューをあらかじめフィルタリングした場合とそうでない場合について，Helpful Review Prediction の精度を比較する実験を行う．ここで，比較文を含むレビューをフィルタリングする手順を以下に示す．

（1） レビューから比較に関連する単語を抽出する
（2） レビューから各単語に対する品詞を取得する
（3） (2) で取得した品詞から比較に関連する品詞を抽出する
（4） (1) または (3) で単語や品詞が抽出できた場合，そのレビューを 「比較文が含まれるレビュー」と定義する

まず (1) において，例えば “than” や “compare”，“prefer” な

ど Jindal ら [1] によって定義された比較に関連する単語が含まれているかどうかを判定する．次に (2) でレビューごとに構文解析を行い各単語に対する品詞を取得する．ここで本研究では，StanfordCoreNLP [10] により構文解析行った．StanfordCoreNLP は，スタンフォード大学が提供している自然言語処理フレームワークであり，比較表現が含まれる品詞を取得できるという特徴を持っている．例えば，他の構文解析ツールでは形容詞を ADJ (形容詞) と品詞付けするのに対し，StanfordCoreNLP は形容詞の中でも JJ (形容詞)，JJR (形容詞:比較級)，JJS (形容詞:最上級) と詳細な品詞付けすることができる．そして (3) において，Jindal ら [1] によって定義された比較に関連する品詞が (2) で取得した品詞に含まれる場合，その品詞を抽出する．最後に，(1) または (3) によって比較に関連する単語や品詞を抽出できた場合，そのレビューを 「比較文が含まれるレビュー」 とする．なお，(1) または (3) によって比較に関連する単語や品詞を抽出できなかった場合，そのレビューは 「比較文を含まないレビュー」 を定義される．

4 実験と結果

本研究では，与えられたレビューに対して *Helpful* であるかどうかを分類する際に，比較文を含んだレビューをあらかじめフィルタリングすることが有効であるかどうかを検討するため，以下の 2 つの分析を行う．1 つ目は，比較文が含まれるレビューと比較文が含まれないレビューでそれぞれ，*Helpful* なレビューと *Unhelpful* なレビューの比率を算出することで，比較文と *Helpful* なレビューの関連性を調査する．2 つ目は，レビューに 「比較文が含まれる」 テストデータセットと 「比較文が含まれない」 テストデータセットを用意し，レビューを *Helpful* か *Unhelpful* かについて 2 値分類を行い，それぞれのテストデータセットの分類結果を比較する．なお，3 節の比較フィルタリング手法に基づいて上記 2 種類のテストデータセットを作成する．

4.1 2 種類のデータセットにおける Helpful と Unhelpful の割合の比較

まず，本節で用いるデータセットについて述べる．Amazon のレビューデータセット [6]のうち，複数の商品名や比較文が多く含まれる傾向にあることから，比較文抽出タスクで使用されることが多く，スマートフォン関連の商品 (Cell Phones and Accessories) についてのレビューを 194,439 件取り出した．その中で，*Helpful* かどうかについて 2 人以上のレビューアによって投票されているレビューである 26,511 件を抽出し，本研究ではこれらのレビューデータを使用する．

これらのレビューデータにおける Helpful Probability (2.2 節参照) の分布図を図 2 に示す．本研究では，Ghose らによって定義された Helpful Probability の閾値に従い，Helpful Probability が 0.6 以上であるレビューは *Helpful*，0.6 未満であるレビューは *Unhelpful* と定義した．その結果，本研究で用いるデータセット 26,111 件のうち，*Helpful* なレビューは 19,393 件，*Unhelpful* なレビューは 7,118 件であった．

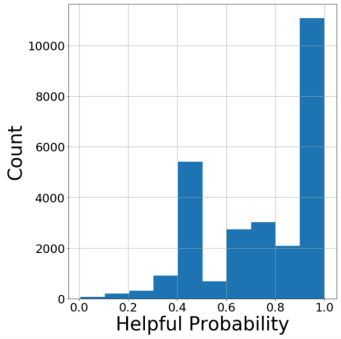

図 2　データセット全体の Helpful Probability の分布図

さらに上記のデータセットを，3 節で述べた比較フィルタリング手法を用いて，比較文を含むレビューのみから構成されるデータセット (以後，比較文ありデータセットと表記する) と比較文を含まないレビューのみから構成されるデータセット (以後，比較文なしデータセットと表記する) に分けた際の，各データセットの内訳を表 1 に示す．表 1 から，比較文ありデータセットは比較文なしデータセットよりも *Helpful* の割合が 7.6% 大きいことがわかる．つまり，本研究で使用したデータセットに関しては，比較文を含まないレビューに比べて比較文を含むレビューは *Helpful* なレビューになりやすいということが言える．さらに，比較文ありデータセットと比較文なしデータセットの間に生じた *Helpful* の割合の差について統計的有意性を確かめるため，2 つの母不良率の違いに関する検定を行った．なお，二項分布の正規近似は永田 [12] の近似法 B1 に従い，統計的検定量 u_0 と p 値を求めた．その結果，$u_0 = 13.721$ と p 値 < 0.001 より，比較文ありデータセットと比較文なしデータセットの間に生じた *Helpful* の割合の差は統計的に有意であるということがわかった．

表 1　*Helpful* と *Unhelpful* の割合

データセット	Helpful	Unhelpful	合計
比較文ありデータセット	8,498 (77.6%)	2,453 (22.4%)	10,950
比較文なしデータセット	10,896 (70.0%)	4,665 (30.0%)	15,561
データセット全体	19,393 (73.2%)	7,118 (26.8%)	26,511

4.2 比較実験

与えられたレビューに対して *Helpful* なレビューであるかどうかを分類する際に，あらかじめ比較フィルタリングを行うことで，*Helpful* なレビューを分類しやすくなると考えられる．そのため本節では，あらかじめ比較フィルタリングを行った場合，Helpful Review Prediction の分類精度が向上するかどうかを

表 2　比較実験に用いるデータセットの内訳

データセット	サイズ
トレーニングデータセット	22,511
比較文テストデータセット	2,000
非比較文テストデータセット	2,000

6 : https://jmcauley.ucsd.edu/data/amazon/

検証する．そこで比較フィルタリングを用いて，レビューに比較文が含まれるテストデータセット (以後，比較文テストデータセットと表記する) と比較文が含まれないテストデータセット (以後，非比較文テストデータセットと表記する) の 2 種類のテストデータセットを用意して Helpful Review Prediction を行った．なお，本研究で使用するレビューデータ全体からこれらのテストデータセットを除いた残りのレビューデータをトレーニングデータセットとした．これらのデータセットのサイズを表 2 に示す．ここで，上記 2 種類のテストデータセットを以下のように設定する．

- 比較文テストデータセット：比較文を含むレビューの中から，2000 件ランダムサンプリングしたレビューデータ
- 非比較文テストデータセット：比較文を含まないレビューの中から，2000 件ランダムサンプリングしたレビューデータ

さらに，各データセットの前処理および本研究で利用する分類モデルについて説明する．まず，品詞分解や単語分解などを行うことができる自然言語処理ツールである NLTK の stopwords list[7] に含まれている stopword をレビューから削除した文を入力とした．そして，本研究では Helpful Review Prediction を行うための分類モデルとして RoBERTa [11] を用いた．RoBERTa は，Wikipedia[8]や OpenWebText[9]などから膨大な文書を事前学習した言語モデルであり，様々な自然言語処理タスクに対して高い精度を達成している．本研究では RoBERTa を用いた分類モデルを実装するにあたり，Simple Transformers[10] というライブラリ内で提供されている事前学習済みの RoBERTa を使用した．なお，本研究で用いた RoBERTa のハイパーパラメータは $learning\,rate = 1e-5$, $epochs = 5$, $batch\,size = 8$, $process\,count = 10$, $seq\,length = 512$ のように設定した．

上記の分類モデルによる分類結果を定量的に評価するための評価指標として，本研究では Accuracy と Precision，Recall，F1–measure を用いた．なお，Positive を *Helpful*，Negative を *Unhelpful* とした 4 つの事象 (True Positive，False Positive，False Negative，True Negative) を用いて，それぞれの評価指標の計算式を以下のように示す．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

- True Positive：正解ラベルが *Helpful* なレビューを分類モデルが *Helpful* であるレビューと予測する事象の数

7：https://gist.github.com/sebleier/554280
8：https://en.wikipedia.org/wiki/Main_Page
9：https://skylion007.github.io/OpenWebTextCorpus/
10：https://github.com/ThilinaRajapakse/simpletransformers

表 3　2 つのテストデータセットの分類結果

	Accuracy	Recall	Precision	F1–measure
比較文テストデータセット	**0.744**	**0.901**	**0.794**	**0.844**
非比較文テストデータセット	0.664	0.847	0.711	0.773

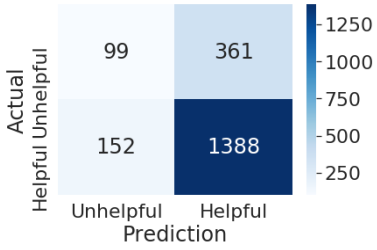

図 3　比較文テストデータセットの分類結果

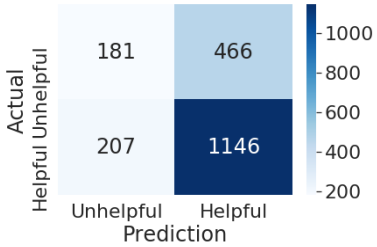

図 4　非比較文テストデータセットの分類結果

- False Positive：正解ラベルが *Unhelpful* なレビューを分類モデルが *Helpful* であるレビューと予測する事象の数
- True Negative：正解ラベルが *Unhelpful* なレビューを分類モデルが *Unhelpful* であるレビューと予測する事象の数
- False Negative：正解ラベルが *Helpful* なレビューを分類モデルが *Unhelpful* であるレビューと予測する事象の数

前述の比較実験の結果を表 3 に示し，各評価指標で優れた結果を太字で示す．表 3 から，全ての評価指標においてあらかじめ比較文でフィルタリングした方が，Helpful Review Prediction の分類精度が優れていることがわかる．また，表 3 の結果より，非比較文テストデータセットより比較文テストデータセットの方が Accuracy において 0.08 ポイント高いことが示されている．この比較文テストデータセットと非比較文テストデータセットの Accuracy の差について統計的有意性を確かめるため，4.1 節と同様に 2 つの母不良率の違いに関する検定を行い，統計的検定量 u_0 と p 値を求めた．その結果，$u_0 = 5.302$ と p 値 $= 3.092e-8$ であり，比較文テストデータセットと非比較文テストデータセットとの Accuracy の差は統計的に有意であるということが示された．

本実験において，分類モデルが比較文テストデータセットと非比較文テストデータセットを用いて分類した結果の Confusion Matrix を図 3 と図 4 に示す．図 3 と図 4 から，比較文テストデータセットと非比較文テストデータセットの間に生じた True Negative と True Positive の差はそれぞれ 82 件と 246 件であることが分かる．つまり，比較文テストデータセットと非比較文テストデータセットにおける True Negative の差よりも True Positive の差の方が多い．したがって，比較文の有無

によって生じる Accuracy の差は True Negative よりも True Positive が強く影響を及ぼしていると言える．そのため本研究では，Accuracy の差に強く影響を与えている True Positive な事象が比較文テストデータセットにおいて多くなる要因について考察する．ここで，実際に比較文を含むレビューが True Positive に該当する例を以下のレビュー例 1〜レビュー例 3 に示す．なお，比較に関連する単語を太字，比較している商品を下線で表す．

レビュー例 1:

The screen is **better than** the Retina display of the iPhone and a gorgeous 4.3in.

レビュー例 2:

Even though the iPhone 5 is 20% **lighter than** any previous iPhone it stills needs protection from falls and to prevent scratching the back.

レビュー例 3:

Although, at $105 at their store, the customer service agent there recommended the Motorola Roadster 2 Universal Bluetooth In-Car Speakerphone because its “Dual Microphone Noise Cancellation and Echo Control setting block out background noise in the car better than other speakerphones” 8.5 DB noise reduction as **compared to** Blueant S4’s 5.0 DB and Jabra Cruiser2’s 3.5 DB.

レビュー例 1 はレビュー対象の商品 (HTC Rezound) と iPhone が 「ディスプレイ」 という観点で比較されており，レビュー例 2 は iPhone5 とそれより前の iPhone が 「軽さ」 という観点で比較されている．そして，レビュー例 3 では，レビュー対象の商品 (Motorola Roadster2) と他の類似商品 (Blueant S4 と Jabra Cruiser2) が 「ノイズ除去」 という観点において比較されている．これらのレビュー例に共通していることは，レビューアが “than” や “compared to” などが含まれる比較文によって特定の観点からレビュー対象の商品とその類似した商品を比較し，優劣をつけているということである．つまり，複数の商品の中でどの観点で対象の商品が優れているのかというユーザにとって価値のある情報量が多く含まれるため，*Helpful* なレビューになりやすいと考えられる．そのため，分類モデルは比較文が含まれるレビューを *Helpful* と分類する傾向が強くなり，True Positive な事象が多くなると考えられる．

一方で，図 3 の比較文テストデータセットの False Positive や False Negative などのように，比較文を含むレビューデータであったとしても分類モデルが誤って分類する場合がある．特に図 3 から分かるように False Negative と比較して False Positive の方が誤った分類を多く行っており，Accuracy の低下に大きな影響を与えていると考えられる．そのため，本研究では False Positive に該当するレビューが多くなってしまう要因について考察する．ここで，実際に比較文を含むレビューが False Positive に該当する例を以下のレビュー例 4 とレビュー例 5 に示す．なお，比較に関連する単語は太字で表す．

レビュー例 4:

I am very happy with the case and the price was great! If you are using a non-OEM charging cable you may need to pop the case off a bit in order to get it to plug in, but **other than** that it has gotten many compliments.

レビュー例 5:

I am pleased with this product. It is **thicker than** I expected (which I like) and of high-quality. It is completely clear. It has detailed instructions and sticky labels that guide you in the application of the product. It has not hindered the sensitivity of the touchscreen in any way. The shape is perfect. I also like that it came with three protectors in one package and has a lifetime replacement warranty. I hope you found this helpful. If I can answer any other questions, please feel free to ask in the comment section. I will reply as quickly as possible.

レビュー例 4 では，“other than” という比較に関連する表現が用いられているが，複数の商品を比較していない．また，レビュー例 5 ではレビュー対象の商品に関して，「厚さ」 という観点で自分の感覚と比べてレビュー対象の商品が厚いと述べているが，レビュー対象の商品に対する比較対象が存在していない．このように，分類モデルが誤分類してしまうレビューにはレビュー対象の商品と比較する商品が存在しないという共通点が見られた．本来，商品の比較を行わない比較文を含むレビューは商品比較を行うレビューよりユーザが得られる情報量が少なくなる．そのため，商品の比較を行わない比較文を含むレビューの正解ラベルは *Helpful* になりにくい傾向がある．その一方，4.1 節で示したように，分類モデルは比較文を含むレビューは *Helpful* になりやすいという傾向を学習しているため，商品比較を行わない比較文を含むレビューに対しても *Helpful* であると誤った予測をしてしまい，False Positive な事象が多くなると考えられる．

さらに，比較文テストデータセットと非比較文テストデータセットの両方において False Positive の事象が多くなってしまう要因を考察する．まず，図 3 と図 4 の Confusion Matrix が示すようにモデルはほとんどのレビューを *Helpful* と予測する結果となった．さらに，比較文テストデータセットは全体の 87.85%，非比較文テストデータセットは全体の 81.15% を *Helpful* と予測している．一方，各テストデータセットで *Helpful* であるレビューは，テストデータセット 1 が全体の 77.0%，テストデータセット 2 は全体の 67.7% であり，どちらのデータセットにおいても正解データより多く *Helpful* と予測していることがわかる．つまり，RoBERTa が *Helpful* に偏った予測をしてしまうため各テストデータセットの False Positive の事象が多くなると考えられる．

次に，False Positive の事象を減少させ，RoBERTa の分類精度を向上させるため，分類モデルが多くのレビューを *Helpful* と予測してしまう要因を考察する．ここで，比較文テストデータセットと非比較文テストデータセット，トレーニングデー

タセットの Helpful Probability の分布図をそれぞれ図 5 と図 6，図 7 に示す．図 7 から Helpful Probability が 1.0 であるレビューは 8,019 件 (トレーニングの 35.6%)，*Helpful* であるか *Unhelpful* であるかの閾値である Helpful Probability が 0.6 以上であるレビューは 16,417 件 (トレーニングデータセットの 72.9%) であり，トレーニングデータセットが偏っていることがわかる．これにより，分類モデルはその偏ったデータを学習し，*Helpful* と予測する傾向が強くなると考えられる．

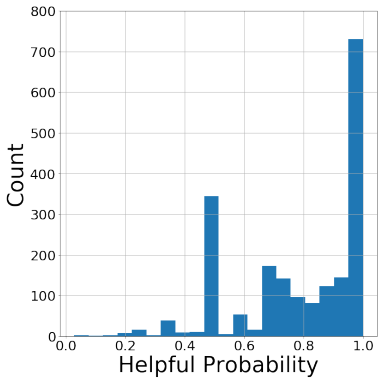

図 5　比較文テストデータセットの Helpful Probability の分布図

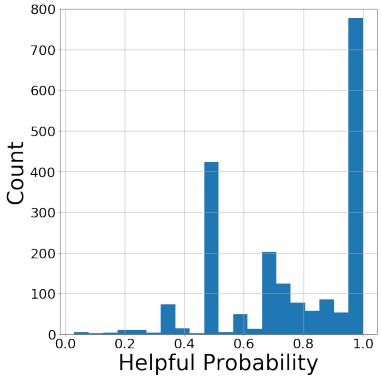

図 6　非比較文テストデータセットの Helpful Probability の分布図

さらに図 5 や図 6 から，トレーニングデータセットだけでなくテストデータセットも偏っていることがわかる．このようにデータセット全体が偏る要因は，Helpful Probability の計算方法にあると考えられる．Helpful Probability は投票した全体の人数とその中で *Helpful* であると投票した人数の割合が同じであれば，たとえ投票した全体の人数が何人であろうと同じ値を取る．例えば，2 人が投票して 2 人が *Helpful* と投票した場合と 100 人が投票して 100 人が *Helpful* と投票した場合はどちらも Helpful Probability が 1.0 となる．つまり，ラベル付けにおける信頼性が異なるにも関わらず同じ Helpful Probality が与えられており，これにより分類モデルが偏った予測を行ってしまう可能性があると考えられる．そのため，*Helpful* であるか *Unhelpful* であるかをラベル付けする際に，投票人数が一定数以上のレビューに対してのみラベル付けを行うなどといった対策を講じるべきであると考えられる．

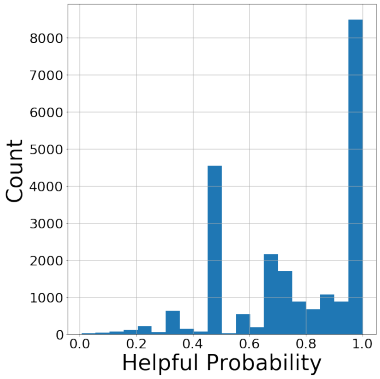

図 7　トレーニングデータセットの Helpful Probability の分布図

5　結論と今後の課題

本研究では，Helpful Review Prediction を行う際に，比較文が含まれるレビューのみをあらかじめフィルタリングすることが有効であるかどうかを検討した．具体的には，レビューに比較文が含まれるテストデータセットとレビューに比較文が含まれないテストデータセットを用意して，Helpful Review Prediction を行い，それらの分類精度を比較する実験を行った．その結果，レビューに比較文が含まれないテストデータセットよりもレビューに比較文が含まれるテストデータセットを分類する方が，優れた分類精度を得ることができた．つまり，本実験で使用したスマートフォン関連の Amazon データセットについて言えば，比較文が含まれるレビューに対してあらかじめフィルタリングを行うと分類モデルがレビューを効率良く分類できると言える．事前に比較フィルタリングすることで商品選定に有用なレビューを疑似的に提示するシステムの構築に役立つと考えられる．一方で，本研究ではスマートフォンに関連するレビューによるデータセットのみを用いたため，本研究の比較フィルタリングが他のドメインでも有効かどうかを検証するため，今後は他のデータセットも利用して同様の評価実験を行う必要があると考えられる．

文　献

[1] Jindal, N. ,Liu, B. “Identifying comparative sentences in text documents. ”SIGIR-06, 2006.

[2] K. Xu, S. Liao, J. Li, Yuxia Song, “Mining comparative opinions from customer reviews for Competitive Intelligence. ” Decision support systems, 50(4), pp. 743–754, 2011.

[3] Yaakov Danone, Tsvi Kuflik, Osnat Mokryn, “Visualizing Reviews Summaries as a Tool for Restaurants Recommendation,” In 23rd International Conference on Intelligent User Interfaces (IUI ’18). ACM, New York, NY, USA, pp. 607–616, 2018.

[4] M.E. Haque, M.E. Tozal, and M. Islam, “Helpfulness Prediction of Online Product Reviews,” In Proceedings of DocEng, pp. 1–4. 2018.

[5] Rudolph Flesch “A new readability yardstick,” Journal of applied psychology, Vol. 32, No. 3, pp. 221. 1948.

[6] S. Mukherjee, K. Popat, G. Weikum, “Exploring Latent Semantic Factors to Find Useful Product Reviews,” SIAM In-

ternational Conference on Data Mining, pp. 480–488. 2017.
[7] Martin L., Pu P., “Prediction of Helpful Reviews Using Emotions Extraction,” AAAI. pp. 1551–1557, 2014.
[8] Malik M., Hussain A., “Helpfulness of product reviews as a function of discrete positive and negative emotions,” Computers in Human Behavior, Volume. 73, 290 - -302,2017.
[9] Ghose A. Ipeirotis, P.G. “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics.” IEEE Trans. Knowl. Data Eng. 23, pp. 1498–1512, 2011.
[10] The Stanford Natural Language Proceeding Groups, Stanford Parser, (https://stanfordnlp.github.io/CoreNLP/), 2018.
[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.
[12] 永田靖, “入門 統計解析法”, pp. 222–225, 日科技連, 2015.