

ポジティブ語・ネガティブ語抽出による
不当なレビュー評価の是正方法に関する一考察

広瀬研究室 4年 C1172313 佐藤利紀

概要

近年、著しい情報技術の発展によって、あらゆる情報が開示され、AIによる作業の効率化がなされている。しかし、どんな作業にも人手が加わり、そこに人為的なミスが生じる。その一例が商品レビューである。商品レビューとはネットショッピングサイトにおいて、その商品に対する評価を評価値とレビュー文で表したものである。商品レビューの集計や整理は自動で行われるものの、商品レビュー自体を記述し投稿する作業は人手で行われる。そのため、評価値とレビュー文の内容が食い違ってしまう可能性がある。たとえば、本の商品レビューにおいて、評価値が低いにもかかわらず「面白い」というレビュー文が書かれているなどの不一致が考えられる。多くのレビューが寄せられた商品の場合、このような評価値と文章が不一致のレビューはあまり問題視されないが、寄せられたレビューが少ない商品の場合、問題が生じる。なぜなら寄せられたレビューが少ない商品の場合、商品レビュー1つがその商品の評価値に及ぼす影響が大きいからだ。その商品レビューの中に評価値と文章が不一致のレビューがあった場合、商品の評価値が正当に反映されず、消費者の購買意思決定の妨げや企業の商品ブランド低下などにつながる恐れがある。そこで本研究では商品カテゴリごとの商品レビューをテキストマイニングし、この問題の解決を図る。具体的には、評価値ごとの主要な単語を見出し、それらの単語を評価値によってその商品カテゴリのネガティブ語とポジティブ語に分類する。これを用い高評価のレビューに含まれるネガティブ語もしくは低評価レビューに含まれるポジティブ語を見出す手法を提案する。そして不当なレビューが混在した商品の評価の是正を目指す。(714字)

目次

第 1 章	はじめに	5
第 2 章	テキストマイニングと関連研究	7
第 3 章	商品レビューの分析	9
3.1	研究手順	9
3.2	商品レビューの収集	9
3.3	RMeCab による形態素解析	10
3.4	ポジティブ語・ネガティブ語の抽出	13
3.5	妥当性の検証	16
3.5.1	商品カテゴリ「本」の高評価レビューにおける有意差の有無	19
3.5.2	商品カテゴリ「本」の低評価レビューにおける有意差の有無	19
3.5.3	商品カテゴリ「ノート PC」の高評価レビューにおける有意差の有無	20
3.5.4	商品カテゴリ「ノート PC」の低評価レビューにおける有意差の有無	20
第 4 章	考察	21
4.1	ポジティブ語・ネガティブ語の抽出に関する考察	21
4.2	不当なレビュー抽出に関する考察	21
第 5 章	結論	25

第1章 はじめに

近年、ネットショッピングの台頭により企業は新たなマーケティングの機会を獲得している [1]。その中で企業が消費者の意見を知る情報源として、商品レビューがある。商品レビューはネットショッピングサイトにおいて、その商品に対する評価を評価値とレビュー文で表したものである。商品レビューを分析することで商品改善や新商品の企画などのマーケティングに活かす試みが行われている [2]。

しかし、商品レビューはレビュワーが人手で記入・投稿するため、意図的な荒らしやヒューマンエラーによる誤評価などにより評価値とレビュー文の内容の食い違いが起きる可能性がある。たとえば、本の商品レビューにおいて、評価値が低いにもかかわらず「面白い」というレビュー文が書かれているなどの不一致が考えられる。多くのレビューが寄せられた商品の場合、このような評価値と文章が不一致のレビューはあまり問題視されないが、寄せられたレビューが少ない商品の場合、問題が生じる。ネットショッピングサイト Amazon¹における商品の評価値は寄せられた商品レビューの評価値の加重平均となっている [3]。これはレビューの多寡によって商品レビュー1つあたりの評価の重みが異なるといえる。そのため、寄せられたレビューが少ない商品だと商品レビュー1つがその商品の評価値に及ぼす影響が大きいと考えられる。その商品レビューの中に評価値と文章が不一致のレビューがあった場合、商品の評価値が正当に反映されず、消費者の購買意思決定の妨げや企業の商品ブランド低下などにつながる恐れがある。そのため現在では E-Guardian²のようなレビューを監視する企業も登場している。しかし、AIで判別が難しい言葉は人手で監視しており、商品レビューのような大量のテキストデータを人手で判断するのは多大な労苦を強いられる。

そこで本研究では商品カテゴリごとの商品レビューをテキストマイニングしこの問題の解決を図る。具体的には、評価値ごとの主要な単語を見出し、それらの単語を評価値によってその商品カテゴリのネガティブ語とポジティブ語に分類する。これを用い高評価のレビューに含まれるネガティブ語もしくは低評価レビューに含まれるポジティブ語を見出す手法を提案する。そして不当なレビューが混在した商品の評価の是正を目指す。

¹<https://www.amazon.co.jp/>

²<https://www.e-guardian.co.jp/>

第2章 テキストマイニングと関連研究

テキストマイニングとは形式化されていない大量のテキストデータを分析して有用な情報を抽出することである [4, 5]。商品レビューのような大量のテキストデータの分析にこのテキストマイニングの技術が用いられる [6]。商品レビュー分析に使われるテキストマイニングでは、主に形態素解析¹によって単語の出現頻度の偏りなどから意見の傾向を可視化する研究がなされている [7]。

商品レビューの分析に関する研究はいくつかある。河野ら [8] の研究では対象、属性、評価表現の単語とその係り受け情報を用いて評価情報を抽出している。那須川ら [9] の研究では「けれども」や「しかし」といった接続表現の周辺に出現する単語を抽出し、好評もしくは不評の極性を付与している。

しかし、河野らの研究では対象、属性、評価表現が完全に揃った文を対象としており、不完全な日本語の表現は想定されていない。たとえば「パソコンの価格は安い」など対象「パソコン」、属性「価格」、評価表現「安い」が揃った文は対象とされているが、「これは安い」といった対象と属性が含まれていない文はこの方法では抽出できない。一方、那須川らの研究では人的過誤による好評、不評の解釈については想定されていない。そこで本研究ではレビューに含まれる単語に焦点をあて、その商品カテゴリ内における高評価、低評価それぞれの特徴となる単語の抽出を試みる。また、不当なレビューに着目した分析、考察を行うことで人為的ミスによって生じる不正確な商品評価の是正を目指す。

¹品詞情報や文法などをもとに文章を単語に分解すること。

第3章 商品レビューの分析

3.1 研究手順

まずネットショッピングサイトの商品レビューを統計分析ソフト R¹を用いて収集する。本研究では評価値と商品カテゴリのフィルタリングが可能なネットショッピングサイト Amazon のレビューを使用する。評価値だけでなく商品化カテゴリごとにも分類する理由は商品カテゴリごとに評価の指標となる単語の偏りが異なると考えたためである。

次に学習用データと評価用データについて述べる。学習用データはレビュー文に含まれる主要な単語の抽出方法を構築するためのレビュー群である。一方、評価用データは構築した抽出方法を評価するためのレビュー群である。収集したレビューをあらかじめこの2つに分類することで後の検証に使用する。

次に主要な単語の抽出方法について述べる。まずレビューの評価値から学習用データの商品レビューを商品カテゴリごとに高評価レビュー、低評価レビューに分類し形態素解析ソフト RMeCab²を用いて形態素解析を行う。次に高評価レビュー、低評価レビューそれぞれの単語の出現頻度の高い順に並べ順位をナンバリングする。ここで高評価、低評価における頻出単語の順位の差を求める。本研究では、この差が大きいほどその商品カテゴリにおける高評価、低評価のレビューの主要な単語とみなし、高評価レビューにおける主要な単語を「ポジティブ語」、低評価レビューにおける主要な単語を「ネガティブ語」と定義する。

最後に評価用データから学習用データで求めたポジティブ語・ネガティブ語を抽出し、その結果の妥当性と評価是正の可能性について考察する。

3.2 商品レビューの収集

Amazon のレビューには以下のような情報が記載されている。

- レビュー文
- 評価値
- レビュー者の名前
- 投稿日時

レビュー者は投稿の際、5段階評価でその商品の評価値を付与する。本研究ではこの評価値が最高の5である商品レビューを高評価レビュー、1もしくは2である商品レビューを低評価レビューとみなした。そして、2020年8月25日時点のAmazonにおける商品カテゴリ「本」と商品カテゴリ「ノートPC」それぞれの商品レビュー合計1600件のレビュー文をRを用いテキストファイル形式で収集した。商品カテゴリ「本」においては、出現単語のばらつきを抑えるため、「ビジネス・経済」ジャンルのみとした。1商品に対して高評価レビュー50件、低評価レビュー50件集め、これを商品カテゴリ「本」においては

¹統計解析に適したフリーソフトウェア。

²Rから日本語のテキストやファイルを指定してMeCabに解析させ、その結果をRで標準的なデータ形式に変換して出力するインターフェイス。

10 商品分、商品カテゴリ「ノート PC」においては6 商品分のレビューを収集した。商品の選定については Amazon 上でその商品カテゴリの売れ筋ランキング上位から、高評価レビュー、低評価レビューを50 件ずつ集めることが可能な商品とした。この1600 件の商品レビューの内、半分を学習用データ、残り半分を評価用データとして活用する(表 3.1)。

表 3.1: 収集した商品レビューの件数の内訳

商品カテゴリ	データ項目	高評価レビュー	低評価レビュー	合計
本	学習用データ	250	250	500
	評価用データ	250	250	500
ノート PC	学習用データ	150	150	300
	評価用データ	150	150	300
	合計	800	800	1600

3.3 RMeCab による形態素解析

次に収集したレビューを RMeCab で形態素解析する。短いテキストを分析する場合は「RMeCabC」コマンド、任意のテキストファイルの形態素解析結果を全て出力する場合は「RMeCabText」コマンドを使用する。その出力結果の一例を以下に示す。

```
> library(RMeCab)
> a <- RMeCabC("とても面白かったです")
> a
[[1]]
  副詞
"とても"

[[2]]
  形容詞
"面白かつ"

[[3]]
  助動詞
"た"

[[4]]
  助動詞
"です"
```

```
> b <- RMeCabText("F:/data/review.txt")
file = F:/data/review.txt
> b
[[1]]
 [1] "とても"      "副詞"        "助詞類接続"
 [4] "*"          "*"           "*"
 [7] "*"          "とても"     "トテモ"
[10] "トテモ"

[[2]]
 [1] "面白かつ"      "形容詞"
 [3] "自立"          "*"
 [5] "*"            "形容詞・アウオ段"
 [7] "連用タ接続"   "面白い"
 [9] "オモシロカツ" "オモシロカツ"

[[3]]
 [1] "た"          "助動詞"   "*"
 [4] "*"          "*"        "特殊・タ"
 [7] "基本形"     "た"       "タ"
[10] "タ"

[[4]]
 [1] "です"         "助動詞"   "*"
 [4] "*"          "*"        "特殊・デス"
 [7] "基本形"     "です"     "デス"
[10] "デス"
```

本研究ではテキストファイルから単語の出現頻度を求めるため、ファイル指定およびテキストの語彙と頻度一覧の作成ができる「RMeCabFreq」コマンドを使用する。そして分析のため解析結果を csv 形式で出力する。csv 形式で出力する流れと、解析結果の冒頭部分を以下に示す³。

```
> book_Hrev_Tra <- RMeCabFreq("F:/review/book_Hrev_Tra.txt")
file = F:/review/book_Hrev_Tra.txt
length = 3765
> write.table( book_Hrev_Tra ,file="F:/review/book_Hrev_Tra_freq.csv",sep=",")
> library(dplyr)
> book_Hrev_Tra %>% head()
      Term      Info1 Info2 Freq
1     その フィラー      *    1
2     なんか フィラー      *    1
3         ま フィラー      *    1
4     ああ 感動詞      *    1
5 ありがとう 感動詞      *    7
6         え 感動詞      *    1
```

この解析を4つの学習用データに行う。そして、得られた解析結果を使いポジティブ語・ネガティブ語の抽出を行う。

³Term は単語、Info1 は品詞大分類、Info2 は品詞細分類、Freq は単語の出現頻度を表す。

3.4 ポジティブ語・ネガティブ語の抽出

次にポジティブ語・ネガティブ語の抽出について述べる。まず、RMeCab を使用し得られた商品カテゴリおよび高評価レビュー、低評価レビューごとの単語の出現頻度からその順位を求める。そして同じ商品カテゴリ内の高評価レビュー、低評価レビューそれぞれにおける単語の出現頻度の順位差を算出した。この順位差を求める式は以下のとおりである。

Hd : 高評価レビューの順位差

Ld : 低評価レビューの順位差

Hr : 当該単語の高評価レビューでの出現頻度順位

Lr : 当該単語の低評価レビューでの出現頻度順位

$$Hd = Lr - Hr$$

$$Ld = Hr - Lr$$

そして各学習データにおいてこれらをまとめた頻出上位 10 単語を表 3.2、表 3.3、表 3.4、表 3.5 に示す。

表 3.2: 商品カテゴリ「本」の高評価レビューにおける頻出上位 10 単語

単語	品詞	品詞細分類	出現頻度	順位	順位差
、	記号	読点	1705	1	0
の	助詞	連体化	1484	2	1
。	記号	句点	1456	3	-1
を	助詞	格助詞	1367	4	4
に	助詞	格助詞	1228	5	4
て	助詞	接続助詞	1196	6	0
は	助詞	係助詞	1120	7	-3
が	助詞	格助詞	981	8	-1
だ	助動詞	*	932	9	-4
する	動詞	自立	897	10	0

表 3.3: 商品カテゴリ「本」の低評価レビューにおける頻出上位 10 単語

単語	品詞	品詞細分類	出現頻度	順位	順位差
、	記号	読点	1661	1	0
。	記号	句点	1384	2	1
の	助詞	連体化	1234	3	-1
は	助詞	係助詞	1086	4	3
だ	助動詞	*	1019	5	4
て	助詞	接続助詞	1014	6	0
が	助詞	格助詞	905	7	1
を	助詞	格助詞	900	8	-4
に	助詞	格助詞	888	9	-4
する	動詞	自立	781	10	0

表 3.4: 商品カテゴリ「ノート PC」の高評価レビューにおける頻出上位 10 単語

単語	品詞	品詞細分類	出現頻度	順位	順位差
。	記号	句点	1216	1	0
、	記号	読点	1112	2	0
た	助動詞	*	1042	3	0
ます	助動詞	*	936	4	4
て	助詞	接続助詞	878	5	0
の	助詞	連体化	833	6	0
は	助詞	係助詞	795	7	2
する	動詞	自立	723	8	-4
に	助詞	格助詞	639	9	1
が	助詞	格助詞	634	10	-3

表 3.5: 商品カテゴリ「ノート PC」の低評価レビューにおける頻出上位 10 単語

単語	品詞	品詞細分類	出現頻度	順位	順位差
。	記号	句点	673	1	0
、	記号	読点	541	2	0
た	助動詞	*	518	3	0
する	動詞	自立	458	4	4
て	助詞	接続助詞	432	5	0
の	助詞	連体化	422	6	0
が	助詞	格助詞	408	7	3
ます	助動詞	*	377	8	-4
は	助詞	係助詞	364	9	-2
に	助詞	格助詞	294	10	-1

表 3.2、表 3.3、表 3.4、表 3.5 から記号、助詞、助動詞の単語が頻出していることが確認できる。またこれらの品詞の単語において、高評価、低評価における単語の出現頻度の順位差は絶対値 5 未満となっており、高評価、低評価の順位差は大きくなりづらいと考えられる。そのため本研究では記号、助詞、助動詞はポジティブ語・ネガティブ語になりづらい単語とみなし、これらの品詞を除外しポジティブ語・ネガティブ語の抽出を試みる。

出現頻度の順位差でポジティブ語・ネガティブ語の抽出を試みるにあたり、単語の有効性の検証のため出現頻度の閾値を設ける必要がある。しかし商品カテゴリごとの学習データのサンプル数が異なるため出現頻度から閾値を求めることはできない。そのため単語の出現頻度を単語の総出現数で割った出現率を求める⁴。本研究ではこの出現率 0.001 を閾値とし、出現率がこの数値以上の単語を有効な単語とみなす。出現率 0.001 以上でフィルタリングし、順位差の降順に並び替えた上位 20 単語を表 3.6、表 3.7、表 3.8、表 3.9 に示す⁵。

⁴小数第六位を四捨五入。

⁵”N/A”は高評価、低評価いずれかのレビュー群に出現しなかったことを表す。

表 3.6: 商品カテゴリ「本」の高評価レビューにおける順位差上位 20 単語

単語	品詞	品詞細分類	出現頻度	出現率	順位	順位差
○	名詞	サ変接続	24	0.00146	96	N/A
分断	名詞	サ変接続	20	0.00122	127	N/A
気づく	動詞	自立	39	0.00238	55	964
見方	名詞	一般	23	0.00140	100	919
教える	動詞	自立	17	0.00104	149	870
本能	名詞	一般	97	0.00591	13	695
+	名詞	サ変接続	54	0.00329	31	677
相手	名詞	一般	43	0.00262	48	660
物事	名詞	一般	21	0.00128	118	590
対人	名詞	一般	19	0.00116	135	573
上司	名詞	一般	17	0.00104	149	559
ひとつ	名詞	副詞可能	17	0.00104	149	559
基づく	動詞	自立	27	0.00164	85	481
すぐ	副詞	助詞類接続	23	0.00140	100	359
事実	名詞	副詞可能	44	0.00268	46	332
悩み	名詞	一般	24	0.00146	96	282
原因	名詞	一般	25	0.00152	92	196
思い込み	名詞	一般	60	0.00366	23	187
勧める	動詞	自立	19	0.00116	135	191
他者	名詞	一般	56	0.00341	28	182

表 3.7: 商品カテゴリ「本」の低評価レビューにおける順位差上位 20 単語

単語	品詞	品詞細分類	出現頻度	出現率	順位	順位差
君	名詞	代名詞	16	0.00112	133	N/A
ページ	名詞	一般	17	0.00119	119	1513
こういう	連体詞	*	15	0.00105	147	1485
人達	名詞	一般	17	0.00119	119	995
使える	動詞	自立	17	0.00119	119	710
彼	名詞	代名詞	17	0.00119	119	710
哲学	名詞	一般	20	0.00140	91	570
なぜ	副詞	助詞類接続	17	0.00119	119	542
設定	名詞	サ変接続	15	0.00105	147	406
アクセス	名詞	サ変接続	22	0.00154	77	400
日本語	名詞	一般	20	0.00140	91	386
印象	名詞	一般	19	0.00133	102	375
チンパンジー	名詞	一般	15	0.00105	147	330
当たり前	名詞	形容動詞語幹	19	0.00133	102	318
語る	動詞	自立	18	0.00126	109	311
無駄	名詞	形容動詞語幹	18	0.00126	109	311
こんな	連体詞	*	18	0.00126	109	311
筆者	名詞	一般	24	0.00168	71	291
トラウマ	名詞	一般	23	0.00161	73	289
感じ	名詞	一般	23	0.00161	73	289

本研究ではこの表 3.6 で示した 20 単語を商品カテゴリ「本」のポジティブ語、表 3.7 で示した 20 単語を商品カテゴリ「本」のネガティブ語とする。同様に表 3.8 で示した 20 単語を商品カテゴリ「ノート PC」のポジティブ語、表 3.9 で示した 20 単語を商品カテゴリ「ノート PC」のネガティブ語とする。

表 3.8: 商品カテゴリ「ノート PC」の高評価レビューにおける順位差上位 20 単語

単語	品詞	品詞細分類	出現頻度	出現率	順位	順位差
自分	名詞	一般	22	0.00175	99	N/A
順	名詞	固有名詞	22	0.00175	99	N/A
ありがとう	感動詞	*	18	0.00144	132	N/A
笑	名詞	固有名詞	16	0.00128	144	N/A
心配	名詞	サ変接続	15	0.00120	152	N/A
最高	名詞	一般	14	0.00112	163	N/A
耐久	名詞	一般	14	0.00112	163	N/A
バージョン	名詞	一般	13	0.00104	173	N/A
機能	名詞	サ変接続	20	0.00159	114	700
買い物	名詞	サ変接続	18	0.00144	132	682
ちょっと	副詞	助詞類接続	17	0.00136	139	675
嬉しい	形容詞	自立	15	0.00120	152	662
速い	形容詞	自立	15	0.00120	152	662
メモリー	名詞	一般	15	0.00120	152	662
2020	名詞	数	15	0.00120	152	662
多少	副詞	助詞類接続	13	0.00104	173	641
安心	名詞	サ変接続	13	0.00104	173	641
軽い	形容詞	自立	24	0.00191	86	439
和	名詞	固有名詞	23	0.00183	92	433
大変	名詞	形容動詞語幹	21	0.00167	104	421

表 3.9: 商品カテゴリ「ノート PC」の低評価レビューにおける順位差上位 20 単語

単語	品詞	品詞細分類	出現頻度	出現率	順位	順位差
返金	名詞	サ変接続	12	0.00190	76	N/A
使い物	名詞	一般	11	0.00174	86	N/A
落ちる	動詞	自立	7	0.00111	156	N/A
広告	名詞	サ変接続	7	0.00111	156	N/A
表記	名詞	サ変接続	8	0.00127	132	1129
失う	動詞	自立	7	0.00111	156	1105
詐欺	名詞	サ変接続	11	0.00174	86	764
最悪	名詞	一般	11	0.00174	86	764
型番	名詞	一般	10	0.00158	104	746
回答	名詞	サ変接続	7	0.00111	156	694
電話	名詞	サ変接続	7	0.00111	156	694
自体	名詞	一般	7	0.00111	156	694
アマゾン	名詞	一般	13	0.00206	71	568
Amazon	名詞	一般	10	0.00158	104	535
仕方	名詞	ナイ形容詞語幹	9	0.00143	116	523
安物	名詞	一般	8	0.00127	132	507
不良	名詞	形容動詞語幹	8	0.00127	132	507
再度	副詞	助詞類接続	7	0.00111	156	483
販売元	名詞	一般	7	0.00111	156	483
ダメ	名詞	形容動詞語幹	7	0.00111	156	483

3.5 妥当性の検証

次に学習用データから抽出したポジティブ語・ネガティブ語の妥当性を評価用データを使って検証する。検証方法としてウィルコクソンの符号付き順位検定を使用する。これは2変量 x および y に関する n 組のデータ $(x_1, y_1), \dots, (x_n, y_n)$ が与えられたとき、それらの差を $z_i = y_i - x_i (i = 1, \dots, n)$ とし、 z_i の母集団分布 $f(z)$ の中央値 θ が0かどうかを調べる手法である [10, 11]。これを利用し、学習用データにおける変量 x と評価用データにおける変量 y との差から、その差の有意性について調べる。変量 x, y については、学習用データおよび評価用データにおけるポジティブ語・ネガティブ語それぞれの出現率を使用する。なぜなら収集したレビューの件数が同じでもそのレビュー内に含まれる単語の総出現数は異なるためだ。そのため学習用データと同様に評価用データの形態素解析を行う。そして前節で選定した各 20 個のポジティブ語・ネガティブ語の出現率を算出した結果を表 3.10、表 3.11、表 3.12、表 3.13 に示す。

表 3.10: 商品カテゴリ「本」の高評価レビューにおけるポジティブ語の出現率

ポジティブ語	品詞	品詞細分類	学習用データにおける ポジティブ語の出現頻度	学習用データにおける ポジティブ語の出現率	評価用データにおける ポジティブ語の出現頻度	評価用データにおける ポジティブ語の出現率
○	名詞	サ変接続	24	0.00146	0	0
分断	名詞	サ変接続	20	0.00122	0	0
気づく	動詞	自立	39	0.00238	34	0.00235
見方	名詞	一般	23	0.00140	2	0.00014
教える	動詞	自立	17	0.00104	24	0.00166
本能	名詞	一般	97	0.00591	1	0.00007
+	名詞	サ変接続	54	0.00329	33	0.00228
相手	名詞	一般	43	0.00262	19	0.00131
物事	名詞	一般	21	0.00128	7	0.00048
対人	名詞	一般	19	0.00116	1	0.00007
上司	名詞	一般	17	0.00104	0	0
ひとつ	名詞	副詞可能	17	0.00104	4	0.00028
基づく	動詞	自立	27	0.00164	1	0.00007
すぐ	副詞	助詞類接続	23	0.00140	20	0.00138
事実	名詞	副詞可能	44	0.00268	11	0.00076
悩み	名詞	一般	24	0.00146	9	0.00062
原因	名詞	一般	25	0.00152	3	0.00021
思い込み	名詞	一般	60	0.00366	1	0.00007
勧める	動詞	自立	19	0.00116	10	0.00069
他者	名詞	一般	56	0.00341	4	0.00028

表 3.11: 商品カテゴリ「本」の低評価レビューにおけるネガティブ語の出現率

ネガティブ語	品詞	品詞細分類	学習用データにおける ネガティブ語の出現頻度	学習用データにおける ネガティブ語の出現率	評価用データにおける ネガティブ語の出現頻度	評価用データにおける ネガティブ語の出現率
君	名詞	代名詞	16	0.00112	0	0
ページ	名詞	一般	17	0.00119	6	0.00052
こういう	連体詞	*	15	0.00105	7	0.00060
人達	名詞	一般	17	0.00119	0	0
使える	動詞	自立	17	0.00119	6	0.00052
彼	名詞	代名詞	17	0.00119	4	0.00034
哲学	名詞	一般	20	0.00140	2	0.00017
なぜ	副詞	助詞類接続	17	0.00119	18	0.00155
設定	名詞	サ変接続	15	0.00105	0	0
アクセス	名詞	サ変接続	22	0.00154	0	0
日本語	名詞	一般	20	0.00140	16	0.00138
印象	名詞	一般	19	0.00133	14	0.00120
チンパンジー	名詞	一般	15	0.00105	0	0
当たり前	名詞	形容動詞語幹	19	0.00133	16	0.00138
語る	動詞	自立	18	0.00126	7	0.00060
無駄	名詞	形容動詞語幹	18	0.00126	19	0.00163
こんな	連体詞	*	18	0.00126	26	0.00223
筆者	名詞	一般	24	0.00168	11	0.00095
トラウマ	名詞	一般	23	0.00161	0	0
感じ	名詞	一般	23	0.00161	14	0.00120

表 3.12: 商品カテゴリ「ノート PC」の高評価レビューにおけるポジティブ語の出現率

ポジティブ語	品詞	品詞細分類	学習用データにおける ポジティブ語の出現頻度	学習用データにおける ポジティブ語の出現率	学習用データにおける ポジティブ語の出現頻度	学習用データにおける ポジティブ語の出現率
自分	名詞	一般	22	0.00175	13	0.00125
順	名詞	固有名詞	22	0.00175	4	0.00038
ありがとう	感動詞	*	18	0.00144	10	0.00096
笑	名詞	固有名詞	16	0.00128	3	0.00029
心配	名詞	サ変接続	15	0.00120	11	0.00106
最高	名詞	一般	14	0.00112	14	0.00135
耐久	名詞	一般	14	0.00112	14	0.00135
バージョン	名詞	一般	13	0.00104	1	0.00010
機能	名詞	サ変接続	20	0.00159	6	0.00058
買い物	名詞	サ変接続	18	0.00144	14	0.00135
ちょっと	副詞	助詞類接続	17	0.00136	13	0.00125
嬉しい	形容詞	自立	15	0.00120	3	0.00029
速い	形容詞	自立	15	0.00120	16	0.00154
メモリー	名詞	一般	15	0.00120	10	0.00096
2020	名詞	数	15	0.00120	8	0.00077
多少	副詞	助詞類接続	13	0.00104	10	0.00096
安心	名詞	サ変接続	13	0.00104	12	0.00115
軽い	形容詞	自立	24	0.00191	11	0.00106
和	名詞	固有名詞	23	0.00183	4	0.00038
大変	名詞	形容動詞語幹	21	0.00167	9	0.00087

表 3.13: 商品カテゴリ「ノート PC」の低評価レビューにおけるネガティブ語の出現率

ネガティブ語	品詞	品詞細分類	学習用データにおける ネガティブ語の出現頻度	学習用データにおける ネガティブ語の出現率	評価用データにおける ネガティブ語の出現頻度	評価用データにおける ネガティブ語の出現率
返金	名詞	サ変接続	12	0.00190	4	0.00083
使い物	名詞	一般	11	0.00174	3	0.00062
落ちる	動詞	自立	7	0.00111	5	0.00104
広告	名詞	サ変接続	7	0.00111	0	0
表記	名詞	サ変接続	8	0.00127	2	0.00042
失う	動詞	自立	7	0.00111	2	0.00042
詐欺	名詞	サ変接続	11	0.00174	5	0.00104
最悪	名詞	一般	11	0.00174	6	0.00125
型番	名詞	一般	10	0.00158	0	0
回答	名詞	サ変接続	7	0.00111	3	0.00062
電話	名詞	サ変接続	7	0.00111	2	0.00042
自体	名詞	一般	7	0.00111	4	0.00083
アマゾン	名詞	一般	13	0.00206	6	0.00125
Amazon	名詞	一般	10	0.00158	7	0.00145
仕方	名詞	ナイ形容詞語幹	9	0.00143	8	0.00166
安物	名詞	一般	8	0.00127	2	0.00042
不良	名詞	形容動詞語幹	8	0.00127	7	0.00145
再度	副詞	助詞類接続	7	0.00111	3	0.00062
販売元	名詞	一般	7	0.00111	1	0.00021
ダメ	名詞	形容動詞語幹	7	0.00111	4	0.00083

$f(z)$ の中央値を θ としたとき、各商品カテゴリおよび各レビューの仮説検定において共通する帰無仮説、対立仮説は以下のとおりである。

帰無仮説 (H_0) : 学習用データ (x) と評価用データ (y) との間に単語の出現率の差はない ($H_0 : \theta = 0$)。

対立仮説 (H_1) : 学習用データ (x) と評価用データ (y) との間に単語の出現率の差はある ($H_1 : \theta \neq 0$)。

本研究では有意水準 0.05 とし仮説検定を行う。検定統計量および P 値の算出には R を使用する。

3.5.1 商品カテゴリ「本」の高評価レビューにおける有意差の有無

表 3.10 で算出した出現率をもとに、商品カテゴリ「本」の高評価レビューにおける学習用データと評価用データで検定を行った結果を以下に示す。

```
Exact Wilcoxon signed rank test

V = 206, p-value = 1.335e-05
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 0.000760 0.001805
sample estimates:
(pseudo)median
 0.0011625
```

P 値が有意水準 0.05 以下であることから帰無仮説が棄却され、対立仮説が採択される。つまり学習用データと評価用データとの間におけるポジティブ語の出現率の差は有意な働きを持つと考えられる。

3.5.2 商品カテゴリ「本」の低評価レビューにおける有意差の有無

表 3.11 で算出した出現率をもとに、商品カテゴリ「本」の低評価レビューにおける学習用データと評価用データで検定を行った結果を以下に示す。

```
Exact Wilcoxon signed rank test

V = 186, p-value = 0.001383
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 0.000285 0.000945
sample estimates:
(pseudo)median
 0.0006225
```

P 値が有意水準 0.05 以下であることから帰無仮説が棄却され、対立仮説が採択される。つまり学習用データと評価用データとの間におけるネガティブ語の出現率の差は有意な働きを持つと考えられる。

3.5.3 商品カテゴリ「ノート PC」の高評価レビューにおける有意差の有無

表 3.12 で算出した出現率をもとに、商品カテゴリ「ノート PC」の高評価レビューにおける学習用データと評価用データで検定を行った結果を以下に示す。

```
Exact Wilcoxon signed rank test

V = 184.5, p-value = 0.001755
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 0.000175 0.000740
sample estimates:
(pseudo)median
 0.000475
```

P 値が有意水準 0.05 以下であることから帰無仮説が棄却され、対立仮説が採択される。つまり学習用データと評価用データとの間におけるポジティブ語の出現率の差は有意な働きを持つと考えられる。

3.5.4 商品カテゴリ「ノート PC」の低評価レビューにおける有意差の有無

表 3.13 で算出した出現率をもとに、商品カテゴリ「ノート PC」の低評価レビューにおける学習用データと評価用データで検定を行った結果を以下に示す。

```
Exact Wilcoxon signed rank test

V = 203, p-value = 3.433e-05
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 0.000385 0.000830
sample estimates:
(pseudo)median
 0.0005975
```

P 値が有意水準 0.05 以下であることから帰無仮説が棄却され、対立仮説が採択される。つまり学習用データと評価用データとの間におけるネガティブ語の出現率の差は有意な働きを持つと考えられる。

表 3.14: ウィルコクソンの符号付き順位検定の結果⁸

商品カテゴリ	レビューの分類	検定統計量	P 値	中央値
本	高評価レビュー	206	0.000013	0.001163
	低評価レビュー	186	0.001383	0.000623
ノート PC	高評価レビュー	184.5	0.001755	0.000475
	低評価レビュー	203	0.000034	0.000598

⁸小数第七位を四捨五入。

第4章 考察

4.1 ポジティブ語・ネガティブ語の抽出に関する考察

ポジティブ語・ネガティブ語の妥当性について、評価用データを使い仮説検定を行った結果、P値がいずれも有意水準である0.05以下であり、妥当性が低い抽出であるという結論となった。これは2つの原因が考えられる。1つはサンプルデータの偏りである。表3.10、表3.11、表3.13において学習用データでは出現したが、評価用データでは一切出現しなかった単語が散見される。これは収集した商品レビューを分類する段階で、学習用データと評価用データとの間において頻出単語が大きく偏っていたと考えられる。そのため、サンプル数を増やし、各データ間の頻出単語の偏りを緩和させることが必要であると考えられる。もう1つの原因は閾値の設定である。本研究では有効な単語を抽出するため、出現率の閾値を設けた。この閾値を適切な値に再設定することで、より正確なポジティブ語・ネガティブ語の抽出が可能であると考えられる。

4.2 不当なレビュー抽出に関する考察

不当なレビュー評価の是正可能性について、評価用データにネガティブ語が含まれる高評価レビュー、ポジティブ語が含まれる低評価レビューがどの程度あるのかを分析した。その結果を表4.1、表4.2、表4.3、表4.4に示す¹。

表 4.1: 商品カテゴリ「本」の高評価レビューにおけるネガティブ語の出現率

ネガティブ語	品詞	品詞細分類	出現頻度	出現率
君	名詞	代名詞	0	0
ページ	名詞	一般	1	0.00007
こういう	連体詞	*	3	0.00021
人達	名詞	一般	0	0
使える	動詞	自立	15	0.00104
彼	名詞	代名詞	10	0.00069
哲学	名詞	一般	1	0.00007
なぜ	副詞	助詞類接続	8	0.00055
設定	名詞	サ変接続	1	0.00007
アクセス	名詞	サ変接続	0	0
日本語	名詞	一般	4	0.00028
印象	名詞	一般	9	0.00062
チンパンジー	名詞	一般	0	0
当たり前	名詞	形容動詞語幹	9	0.00062
語る	動詞	自立	3	0.00021
無駄	名詞	形容動詞語幹	3	0.00021
こんな	連体詞	*	5	0.00035
筆者	名詞	一般	5	0.00035
トラウマ	名詞	一般	0	0
感じ	名詞	一般	7	0.00048

¹出現率は小数第六位を四捨五入。

表 4.2: 商品カテゴリ「本」の高評価レビューにおけるポジティブ語の出現率

ポジティブ語	品詞	品詞細分類	出現頻度	出現率
○	名詞	サ変接続	0	0
分断	名詞	サ変接続	0	0
気づく	動詞	自立	4	0.00034
見方	名詞	一般	0	0
教える	動詞	自立	5	0.00043
本能	名詞	一般	2	0.00017
+	名詞	サ変接続	10	0.00086
相手	名詞	一般	7	0.00060
物事	名詞	一般	0	0
対人	名詞	一般	0	0
上司	名詞	一般	1	0.00009
ひとつ	名詞	副詞可能	0	0
基づく	動詞	自立	1	0.00009
すぐ	副詞	助詞類接続	6	0.00052
事実	名詞	副詞可能	10	0.00086
悩み	名詞	一般	3	0.00026
原因	名詞	一般	3	0.00026
思い込み	名詞	一般	0	0
勧める	動詞	自立	9	0.00077
他者	名詞	一般	1	0.00009

表 4.3: 商品カテゴリ「ノート PC」の高評価レビューにおけるネガティブ語の出現率

ネガティブ語	品詞	品詞細分類	出現頻度	出現率
返金	名詞	サ変接続	2	0.00019
使い物	名詞	一般	0	0
落ちる	動詞	自立	2	0.00019
広告	名詞	サ変接続	4	0.00038
表記	名詞	サ変接続	1	0.00010
失う	動詞	自立	0	0
詐欺	名詞	サ変接続	0	0
最悪	名詞	一般	0	0
型番	名詞	一般	0	0
回答	名詞	サ変接続	0	0
電話	名詞	サ変接続	1	0.00010
自体	名詞	一般	9	0.00087
アマゾン	名詞	一般	0	0
Amazon	名詞	一般	4	0.00038
仕方	名詞	ナイ形容詞語幹	7	0.00067
安物	名詞	一般	3	0.00029
不良	名詞	形容動詞語幹	6	0.00058
再度	副詞	助詞類接続	5	0.00048
販売元	名詞	一般	0	0
ダメ	名詞	形容動詞語幹	4	0.00038

表 4.4: 商品カテゴリ「ノート PC」の低評価レビューにおけるポジティブ語の出現率

ポジティブ語	品詞	品詞細分類	出現頻度	出現率
自分	名詞	一般	10	0.00208
順	名詞	固有名詞	1	0.00021
ありがとう	感動詞	*	0	0
笑	名詞	固有名詞	0	0
心配	名詞	サ変接続	5	0.00104
最高	名詞	一般	1	0.00021
耐久	名詞	一般	1	0.00021
バージョン	名詞	一般	3	0.00062
機能	名詞	サ変接続	16	0.00332
買い物	名詞	サ変接続	2	0.00042
ちょっと	副詞	助詞類接続	9	0.00187
嬉しい	形容詞	自立	1	0.00021
速い	形容詞	自立	2	0.00042
メモリー	名詞	一般	3	0.00062
2020	名詞	数	0	0
多少	副詞	助詞類接続	1	0.00021
安心	名詞	サ変接続	0	0
軽い	形容詞	自立	3	0.00062
和	名詞	固有名詞	1	0.00021
大変	名詞	形容動詞語幹	3	0.00062

表 4.1、表 4.3 を参考に、ネガティブ語が高評価レビューにどのような文脈で含まれているのか分析した結果、大きく 3 つのパタンがあると考えられる。1 つ目は文脈によってポジティブ・ネガティブの極性が変化するパタンである。その例として、以下のようなネガティブ語の出現パタンがみられた²。

- ……こういう方法があったのか、と衝撃を受けました。……
- ……金額返金保証……
- ……再度購入しました。……

上記のようにネガティブ語がポジティブな表現としてレビュー文に含まれている。そのため文章全体においてその単語がどのような文脈で用いられているのかをポジティブ語・ネガティブ語の抽出に組み込み、両方の極性を持つ単語を除外するなどの処理が必要であると考えられる。

2 つ目は評価対象の商品とは無関係の文脈に現れるパタンである。その一例を以下に示す。

- ……日本語教師という仕事柄、……
- ……前に他社メーカーのものを買ったが初期不良が多かったため、……
- ……がダメになり買い替えました。……

上記から自分の身の上話や以前購入していた商品に対する言及などレビューのエピソードに関する文脈に出現していることが確認できる。そのため言及している対象が明確な場合、その対象が評価対象の商品とどの程度関連性が高いかを考慮する必要がある。また、身の上話のような評価対象の商品と関連性の低い文章を除外するため、単語の共起関係を分析する必要があると考える。

3 つ目は逆説表現の文脈に現れるパタンである。その一例を以下に示す。

- ……読めば当たり前に感じるかもしれませんが、……
- ……PC自体古いとはいえ……
- ……質感は安物感が半端ないですがコストパフォーマンスが抜群……

²レビュー文の一部を抜粋。下線は本研究におけるポジティブ語・ネガティブ語を表す。

前記から「が」や「とはいえ」などの逆説表現の接続詞が含まれていることが確認できる。そのためポジティブ語・ネガティブ語の抽出において、このような文脈の可能性も考慮する必要がある。しかしこの逆説表現の接続詞に焦点を当てることで高評価レビューに含まれるネガティブな表現もしくは低評価レビューに含まれるポジティブ表現が抽出が可能となり、不当なレビュー発見につながると考える。

一方、表 4.2、表 4.4 を参考にポジティブ語が低評価レビューにどのような文脈で含まれているのか分析した結果、前述のパタンに加え、否定表現の文脈に現れるパタンがみられた。その一例を以下に示す。

- ……全体として気づきもなく、……
- ……お勧めできません。……

上記のようにポジティブ語の後に「ない」、「できない」などの前の単語を否定する表現が付随しており、ポジティブ語がネガティブな表現となっている。そのため文章内に出現するポジティブ語・ネガティブ語の周囲における否定表現の有無を調べる必要がある。

第5章 結論

本研究では不当なレビュー評価是正を目指し、単語の出現頻度の順位差を利用した商品レビューからポジティブ語・ネガティブ語の抽出を試みた。そしてウィルコクソンの符号付き順位検定を行い、その商品カテゴリ内のポジティブ語・ネガティブ語が妥当なものであるかを明らかにした。これにより商品レビューのテキストマイニングによるポジティブ語・ネガティブ語抽出の可能性を見出した。

今回は商品レビュー最大 250 件を学習用データとしポジティブ語・ネガティブ語の抽出を行ったが、評価用データと異なる抽出結果となり、各データにおける出現単語の偏りがみられた。そのためサンプル数を増やし、抽出するポジティブ語・ネガティブ語の妥当性を高める必要がある。また実際に不当なレビューを抽出するためには、レビュー文内における単語同士の関係性を分析する必要がある。複数のテキストマイニングの手法を併用し、ポジティブ語・ネガティブ語の抽出および抽出単語の検証をすることが今後の課題としてあげられる。

参考文献

- [1] 金鍾和, 森高正博, 福田晋, 尹ソク重. ネットショッピングにおける消費者購買認識の構造分析. フードシステム研究第 19 巻 4 号, pp.382-393, 2013
- [2] 佐藤利紀. 品詞共起パターンを用いた評価表現抽出方法. 会津大学短期大学部産業情報学科経営情報コース 2018 年度卒業論文要旨集, pp.41-44, 2019
- [3] Ben Fox Rubln, cnet.”<https://www.cnet.com/news/amazon-updates-customer-reviews-with-new-machine-learning-platform/>”. 2015
- [4] 石田基広. R によるテキストマイニング入門 第 2 版, 森北出版, 2017
- [5] 小林雄一郎. R によるやさしいテキストマイニング, オーム社, 2017
- [6] 那須川哲哉. テキストマイニングを使う技術/作る技術:基礎技術と適用事例から導く本質と活用法, 東京電気大学出版局, 2006
- [7] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, 13(3), pp.201-241, 2006
- [8] 河野一志, 町田翔, 村松拓実, 小林美穂, 延澤志保. 不特定分野のレビューを対象とした評価情報の自動認識. 情報処理学会第 79 回全国大会, 3Q-06, pp.535-536, 2017
- [9] 那須川哲哉, 金澤博. 文脈一貫性を利用した極性付評価表現の語彙獲得. 情報処理学会研究報告自然言語処理, pp.109-116, 2004
- [10] 岩崎学. 統計的データ解析入門 ノンパラメトリック法, 東京図書, 2006
- [11] 村上秀俊. 統計解析スタンダード ノンパラメトリック法, 朝倉書店, 2015
- [12] 吉見憲二. グルメサイトの信頼性確保に関する一考察. 情報処理学会研究報告マルチメディア通信と分散処理研究会報告 2014-DPS-161(2), pp.1-4, 2014
- [13] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌, 47(2), pp.627-637, 2006
- [14] 小倉達也, 宍戸開, 今藤紀子, 山口実靖, 浅谷耕一. レビューサイトにおける良質なレビューの特性とそれを考慮した評判情報の抽出に関する一考察. DEWS2008, B8-5, 2008
- [15] 清水航. レビューの語の重みを考慮したテキストマイニングによるゲームソフトの評判分析. 法政大学大学院工学研究科 2010 年度修士論文. 2011
- [16] 倉橋宏幸, 青野雅樹. Amazon レビューを用いた有用性の判別実験. 情報科学技術フォーラム講演論文誌, 12(2), pp.101-102, 2013