Journal of Data Intelligence, Vol. 1, No. 4 (2020) 468–488
© Rinton Press

CLAUSE-LEVEL ANALYSIS HIGH-VALUE REVIEWS BASED ON SENTIMENT

AKIYO NADAMOTO
Konan University, 8-9-1 Okamoto Higashinada-ku
Kobe, Japan
nadamoto@konan-u.ac.jp

KAZUHIRO AKIYAMA
Konan University, 8-9-1 Okamoto Higashinada-ku
Kobe, Japan
f4115um@gmail.com

TADAHIKO KUMAMOTO
Chiba Institute of Technology, 2–17–1, Tsudanuma
Narashino, Japan
kumamoto@net.it-chiba.ac.jp

Today, huge numbers of reviews are posted on the internet. Online shoppers often refer to reviews written about the products. A review has a star rating that represents what other people think about the product. However, the star rating is not always appropriate for evaluating the product. High-value reviews that affect the users' willingness to buy are independent of the number of stars in ratings. High-value reviews are those from which people find useful information those regarded as good reviews. As described in this paper, we investigated the relation between high-value reviews and the sentiment (positive/negative/neutral) of their clauses based on four hypotheses. We extract characteristics of high-value reviews based on our results. Furthermore, we propose a classification method that classifies clause level sentiment from reviews.

Keywords: Review Sentiment CRF Clause

1. Introduction

Along with the rapid progress of E-commerce websites, people often buy products on the internet. Actually, E-commerce websites offer so many products that it is difficult for a person to find their preferred product. Therefore, when people practice online shopping, they often refer to reviews that are written about the products. They can understand the products more deeply by reading the reviews. Moreover, they use the reviews as criteria for buying products. A White Paper on Information and Communications published by the Ministry of Internal Affairs and Communications of Japan[1] reports that 7080% of people who are 2060 years old refer to reviews when they use online shopping. As described herein, a person who reads reviews is a "user". Therefore, users regard reviews as important when using online shopping. However, the number of reviews becomes huge. Certainly, not all reviews are useful to users. For example, when a user wants to buy a lightweight laptop computer with high-performance through online shopping, the user reads reviews that are written about laptop computers. Some reviews are helpful for users, but many reviews might include useless or nuisance information. Therefore, finding reviews that include useful information is difficult.

Table 1. Examples of high-value reviews

(1)	I used various types of toilet paper, but nowadays I always use this toilet paper.
(2)	I use this toilet paper for a long time because it is twice as long as usual.
(3)	The toilet paper is very soft, but the size is too big to fit in the holder.
(4)	I like the toilet paper because of its softness and because the smell is not too strong.

Moreover, reviews typically have a star rating showing what other people think about the product. People sometimes refer specifically to the star rating when they read the reviews. However, the star rating is not always appropriate for evaluating the product. Figure 1 portrays the relation between the star rating and the ratio of users who push the “helpful” button on Amazon. Herein, we designate the number of users who push the “helpful” button on Amazon as “users’ evaluations”. The vertical axis is the ratio of the number of all reviews assigning each star level to the number of reviews per rating. From Figure 1, one can infer that the star rating number is not related to the number of users’ evaluations. The number of stars rating is unrelated to high-value reviews.

Ando et al.[2] analyzed the reviews and proposed that reviews with product evaluation and sentiment affect the users’ willingness to buy. Especially, they reported that such reviews include product specifications and good points and bad points supporting their sentiments related to products. As described herein, we designate such a review that affects users’ willingness to buy as “high-value reviews”. Our proposed high-value review is independent of the number of star ratings. A high-value review is one by which users find useful information when they read the review. They regard it as a good review.

Table 1 presents examples of high-value reviews. We can also generalize some characteristics of high-value reviews from the examples. High-value reviews include (1)a simple summary of the usability and (2)good points and bad points of the products.

A review of product usability shows that high-value reviews consist of classifiable sentences: neutral sentences; positive sentences presenting good points about the products; and negative sentences presenting bad points about products. Investigating review sentiments can be helpful to find high-value reviews easily. As described in this paper, we investigate relations between high-value reviews and their sentiments. Our analysis of sentiment has three axes: positive, negative, and neutral. When investigating the relation between high-value reviews and their sentiment, we generate four hypotheses and discuss the results individually. For this research, we can contribute that, based on analyses of high-value reviews, we can extract documents which readily affect a user’s willingness to buy.

Furthermore, from our investigations, we can understand that high-value reviews have multiple sentiments. However, automatic determination of positive and negative points in a review is difficult. Nevertheless, many studies have undertaken sentiment classification. Their target is long sentence(s) or documents. However, sentences in a review often consist of multiple sentiments, even if they are short sentences. In this paper, we propose sentiment classification using the clause unit as the minimum unit of sentiment classification of a review. Specifically, we use the Conditional Random Field (CRF), which is a method to solve series labeling by structural learning for sentiment estimation of each clause, considering not only the clause by which sentiment classification is performed, but also its relation with its neighboring clause.

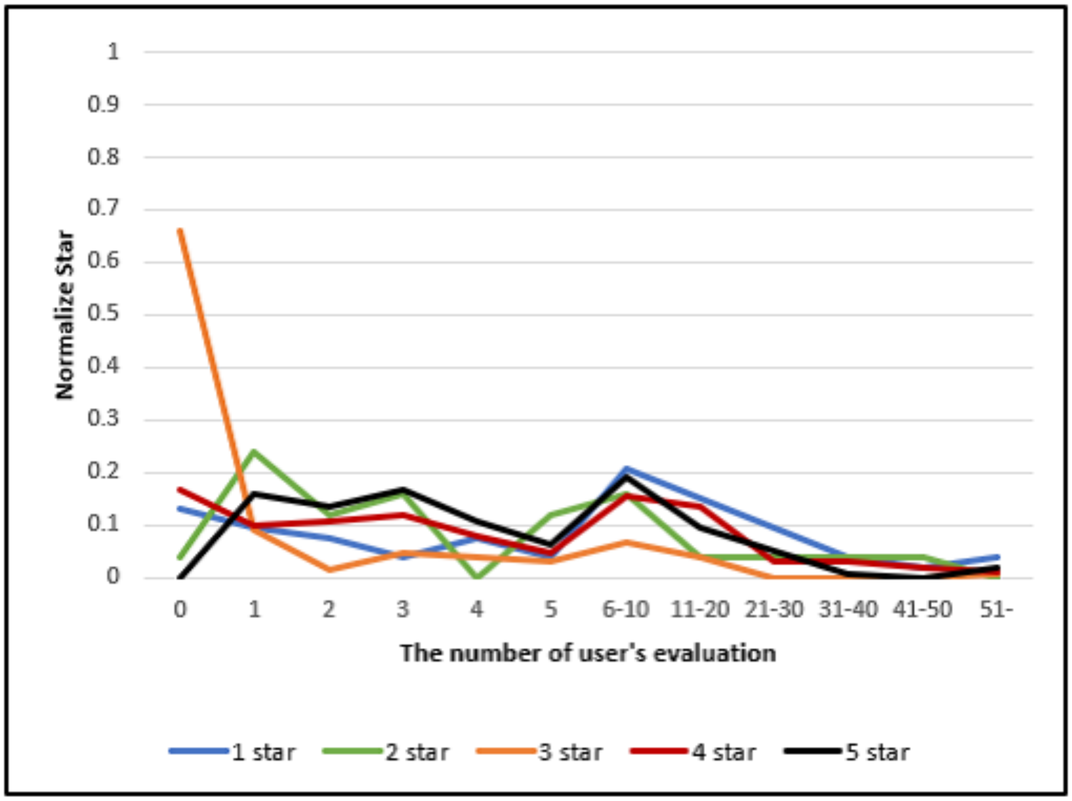

Fig. 1. Star rating and the ratio of users' evaluations.

This paper is organized as follows: Section 2 presents a discussion of related work. Section 3 presents an analysis of high-value reviews based on the sentiment. Section 4 proposes the clause-level sentiment classification using CRF. Finally, Section 5 presents a description of the paper conclusions.

2. RELATED WORK

Analysis reviews

Many studies have examined extraction of good reviews and analysis of many review types. Wu et al.[3] present ReviewMiner, an aspect-based sentiment classification system for entities in an online review. The system summarizes and visualizes opinions to provide rich perspectives to users. Chatterjee et al.[4] investigate effects of negative reviews such as product and company complaints. They also discuss strategies to reduce effects of negative word of mouth. Nanli et al.[5] present surveys of the latest development in sentiment analysis of reviews, and make in-depth introductions of research and applications in business and the Blogsphere. Yohan et al.[6] tackle the problem of automatically discovering what aspects are evaluated in reviews and how sentiments for different aspects are expressed. Furthermore, we consider evaluation and results. Devika et al.[7] use sentiment analysis for numerous reviews with various approaches to analyze and compare them. Results show that information necessary for customers is extracted, and show its power. Almost all of these studies examine sentence-level or document-level sentiment. By contrast, we specifically examine clause-level sentiment.

Ly et al.[8] not only extract review sentiments but also extract the underlying justification for their opinion. Furthermore, they achieve this through novel application of clustering and validate our approach through empirical study. C. Liu et al.[9] design and develop a movie-rating and review-summarization system. The movie-rating information is based on the sentiment-classification result. The condensed descriptions of movie reviews are generated from feature-based summarization. Li et al.[10] study automatic review mining and summarization. This study extracts features based on which the reviewers express their opinions and determine whether the opinions are positive or negative. Abulaish et al.[11] present an opinion mining system to identify product features and opinions from review documents. They are extracted using semantic and linguistic analyses of text documents. Liu et al.[12] address difficulties of detecting low-quality product reviews. To extract low quality product reviews, they use two methods. Fangtao et al.[13] specifically examine object feature based review summarization. They use Conditional Random Fields (CRFs) for a new machine learning framework. Additionally, they verify its accuracy. These studies specifically examine review contents and generate reviews that are easier to understand. However, for the present study, we examine targets of multiple product types of reviews and assess differences between review ratings and product types.

Sentiment classification

Recently, studies of sentiment analysis have been actively pursued. Especially, assessing the sentiments of reviews is important to extract user opinions. Sudhof et al.[14] develop a theory of conditional dependency between emotional states. They are characterized not only by polarity and intensity but also by the role they play in state transitions and social relationships. Zhang et al.[15] identify sentence-level sentiments based on sentence structure and context information. They propose a conditional random field method with two active learning strategies for labeling sentiment data. Zhao et al.[16] propose a CRF-based method that responds to two special characteristics of "contextual dependency" and "label redundancy". They introduce redundant labels into the original sentimental label set to capture the contextual constraints on sentence sentiment. Patra et al.[17] develop an aspect-based sentiment classification system using CRF. Their system identifies aspect terms, aspect categories, and their sentiments from review sentences. Yang et al.[18] assess sentence-level sentiment of reviews. They applied posterior regularization to predict CRF parameters. Choi et al.[19] present a novel learning approach to ascertain the polarity. They consider interaction among words such as negator. Rustamov et al.[20] characterize each sentence of the customer review in polarity and subjectivity for sentiment classification. Socher et al.[21] propose Recursive Neural Tensor Network (RNTN), the improvement model of Recursive Neural Models (RNN, MV-RNN). They classify movie reviews as positive, negative, or neutral at the sentence and document levels. Almost all of these studies examine sentence or document-level sentiment. By contrast, we specifically examine clause-level sentiment classification.

3. Analyzing relations between high-value reviews and their sentiments

When investigating the relations between high-value reviews and their related sentiments, we generate four hypotheses and discuss the results individually.

3.1. *Conditions of Analysis*

(a) Flow of analysis

The steps of investigating relations between high-value reviews and their sentiments are the following:

1. Collect reviews by product type automatically and randomly.
2. Divide a review into clauses.
3. Ascertain the sentiments of each clause using crowdsourcing.
4. Analyze relations between high-value reviews and their sentiment based on our four proposed hypotheses.

(b) Datasets

We analyze reviews for products of five types from three Amazon categories: home&kitchen, electronics, and cosmetics. We analyze reviews of toilet paper and insect repellent from the home&kitchen category, laptop computers and smartphones from the electronics category, and makeup base from the cosmetics category. We collected 500 reviews for each product. The total reviews are 2,500.

(c) Dividing a review into clauses

Some reviews present multiple sentiments. For example, the review of “This vacuum cleaner performs well but makes noise.” includes multiple sentiments. “This vacuum cleaner performs well” is a positive sentiment but “makes noise.” is a negative sentiment. In this way, we correct sentiments that can not be analyzed on a sentence or a review level. Furthermore, when we analyze the sentiment based on a word or phrase, the semantics of the review is lost. Therefore, we analyze reviews based on a clause.

When we divide each review into clauses, we use the Japanese Dependency and Case Structure Analyzer KNP created by Kurohashi et al. [22]. First, we conduct morphological analysis of reviews using Juman (a User-Extensible Morphological Analyzer for Japanese)[a]. Next, we apply a parser using KNP. Then we divide a review into multiple clauses. For example, the review of “I love this detergent because it cleans well and it is good for my hands.” is divided into three clauses, which are “I love this detergent.”, “because it cleans well.”, and “and it is good for my hands.”.

(d) Determining the sentiment of each clause

We determined the sentiment of each clause using crowdsourcing. Of the crowd, 10 workers tagged a sentiment to a clause. They read the clause and judged whether the clause is positive, negative or neutral. We regard the clause as positive when it was judged as positive by more than 6 out of 10 workers. We regard negative and neutral similarly. Moreover, when

[a] JUMAN http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

six or more workers do not judge a single sentiment, we regard the clause as neutral. Table 2 presents results of sentiment determination. Results of sentiment determiniation show moderate agreement because the kappa coefficient is 0.40.6.

Table 2. Results of sentiment evaluation by crowdsourcing

Data			Number of clauses			kappa coefficient
domain	category	product	Positive	Negative	Neutral	
Amazon	Daily necessities	Toilet paper	955 (31%)	442 (14%)	1721 (55%)	0.473
		Insect repellant	777 (24%)	370 (12%)	2492 (64%)	0.460
Amazon	Home appliance	Laptop	1054 (21%)	858 (17%)	3187 (62%)	0.464
		Smartphone	1118 (20%)	1086 (19%)	3505 (61%)	0.513
Amazon	Cosmetic	Makeup base	1122 (34%)	631 (19%)	1582 (47%)	0.529

(e) Analyzing relations between high-value reviews and their sentiment

- Conditions of high-value reviews
  Conditions of high-value reviews are the following:
  - The condition is not dependent on the number of star ratings.
  - The condition includes information that users regard as useful and which they regard as a good review.

  Then we calculate our dataset based on users' evaluations of reviews (see Table 3). From the results, when analyzing the review sentiment, we regard high-value reviews as those regarded by more than three users as a high-value review.
- Review sentiment
  When analyzing the relation, we use sentiment across reviews, combining the sentiment of each clause. Therefore, the number of sentiments included in a review is equal to the number of clauses.

We analyze the relation between high-value reviews and their sentiment based on our proposed four hypotheses.

3.2. *Hypothesis*

When we analyze sentiments of high-value reviews, we propose the following four hypotheses.

Hypothesis 1. High-value reviews have both positive and negative clauses.
When a review has only either positive or negative clauses, except for neutral clauses, the author writes only good point(s) or bad point(s) for the product. A high probability exists that the review is biased by the author because of love for the product, an assumption, or

Table 3. Number of users' evaluations for our dataset reviews

# Users' evaluation	Product				
	Toilet paper	Insect repellant	Laptop	Smartphone	Makeup base
0	213	213	61	130	130
1	125	92	80	65	122
2	58	63	61	42	54
3	27	33	45	51	49
4	23	30	38	35	26
5	13	13	29	24	23
6-10	34	41	73	71	56
11-20	7	13	57	43	23
21-30	0	2	27	17	8
31-40	0	0	11	11	1
41-50	0	0	7	4	1
51-	0	0	11	7	7

deception. However, when a review has both positive and negative clauses, the author includes both good point(s) and bad point(s) in the review. The author presumably is writing a review from a fair perspective. Users usually want to know good point(s) and bad point(s) of a product when they buy it. Therefore, we propose hypothesis 1. In this analysis, "both positive and negative clauses" signifies that one or more positive and one or more negative clauses are included in a review.

Hypothesis 2. High-value reviews include many neutral clauses.
We consider that a review of the mention of usability and features of a product is a neutral clause(s). In other words, a high probability exists that a review with many neutral clauses is written to explain the usability and features of a product. Users want to know the usability and features of the product from a review. Therefore, a review with many neutral sentiments is regarded as a high-value review. Then we propose hypothesis 2. In this analysis, we regard a review in which three neutral clauses are consecutive as a review with many neutral clause reviews.

Hypothesis 3. The trend of high-value review is the same in different product categories.
Review words vary by product category. However, we believe that sentiments of high-value reviews are the same across different product categories because users want to know good points and bad points of a product and because their sentiments are the same across multiple product categories. Therefore, we propose hypothesis 3.

Hypothesis 4. Reviews by expert reviewers have many neutral clauses.
Expert reviewers who have high-level knowledge of the product write with specific citations such as features of the products and usability of the products. The information includes many neutral clauses. Therefore, we consider reviews by expert reviewers as having many neutral clauses. We propose hypothesis 4.

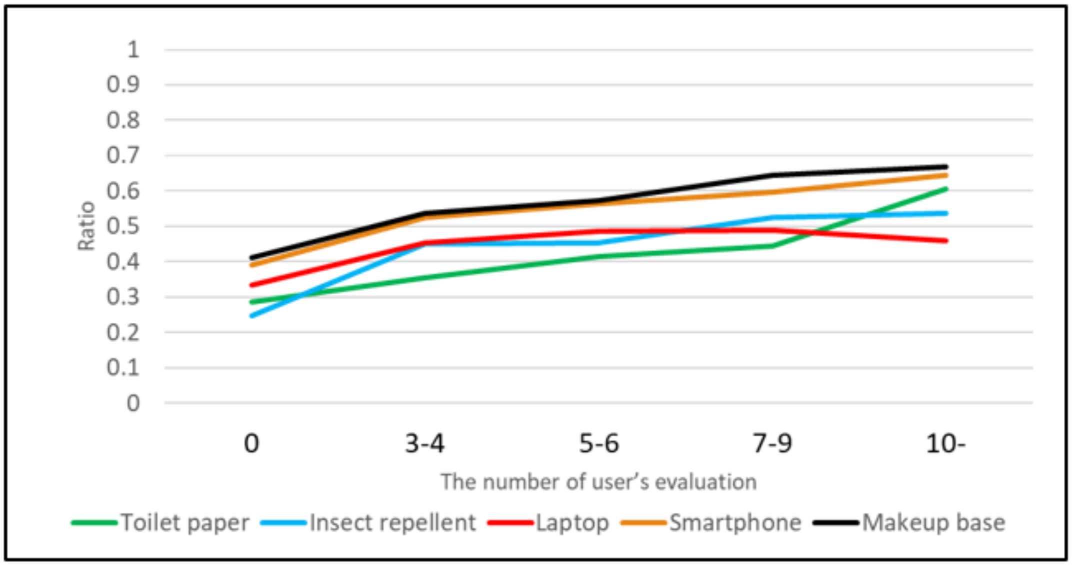

Fig. 2. Rate of positive and negative clauses.

3.3. *Results and Discussion*

Hypothesis 1. High-value reviews include both positive and negative clauses. Table 4 and Figure 2 show results: the reviews with both positive and negative clauses and the number of users' evaluations. Table 5 and Figure 3 present results which are the rate of the reviews with only positive clauses and the number of users' evaluations. Table 6 and Figure 4 show results which are the rate of the reviews with only negative clauses and the number of users' evaluations. No result includes consideration of neutral clauses. Rate RR_i is the following.

$$RR_i = \frac{SC_i}{\sum_{j=1}^{k}(NC_j)}$$

Therein, i denotes the types of clause sentiments which are positive&negative, positive only, and negative only. Also, SC_i denotes the number of i sentiment clauses in users' evaluation clauses for each number of users. The number of reviews in each number of users is k. Also, NC_j stands for the total number of clauses in a review for each number of users. From Figure 2, one can understand that users regard reviews that include both positive and negative clauses as high-value reviews because the graph is increasing. Furthermore, in Figure 3, the graph is decreasing. We can understand that users do not regard such reviews, which consist only of positive clauses, as high-value reviews. However, in Figure 4, the graph is slightly increasing, from which one can infer that users draw on such negative clauses.

From these results, hypothesis 1 is verified as "True". Moreover, users draw on negative clauses.

Hypothesis 2. High-value reviews have many neutral clauses. Table 7 and Figure 5 show reviews with many neutral clauses. The graph is increasing. How-

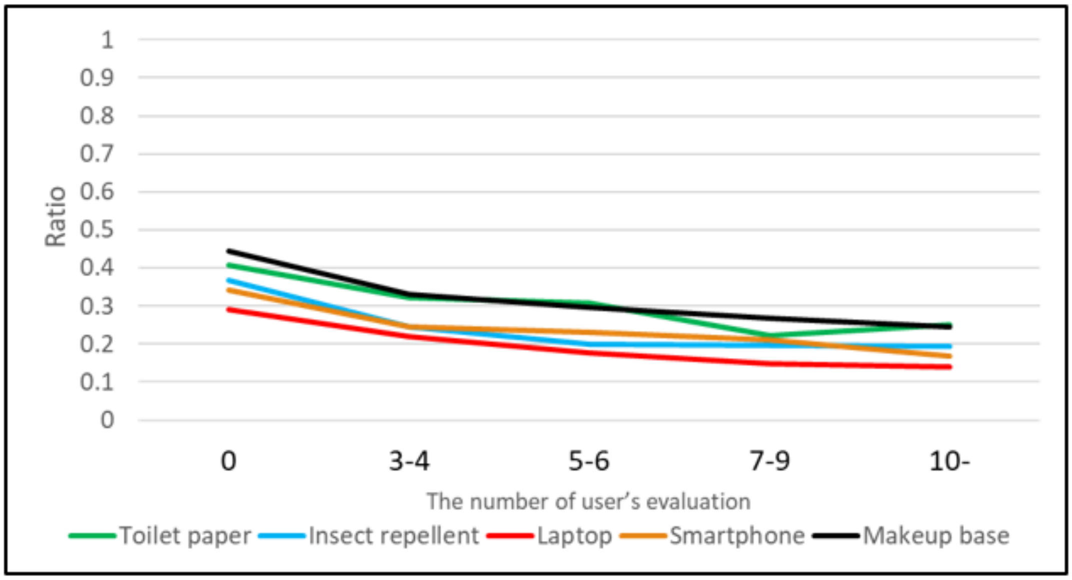

Fig. 3. Rate of positive clauses.

ever, the graph becomes flat when the number of users' evaluations becomes greater than 7. Results for Hypothesis 1 and Graph 3 show that we can infer that a review with many users is regarded as a high-value review, including not only neutral clauses but also positive and negative clauses. Therefore, high-value reviews include not only usability and product features but also good points and bad points of the products. We can understand hypothesis 2 as "True". Moreover, additional information of hypothesis 2 is that high-value reviews include positive, negative, and neutral clauses.

Hypothesis 3. Trends of high-value reviews are the same in different product categories.
In Table 4, almost all products increase in parallel, except that of laptops. The result for laptop computers is flat in the graph. The reason is that the contents of reviews about laptop computers differ between those posted by reviewers who are experts on laptops and those who are beginners at using laptop computers. As described in this paper, we designate reviewers who are experts at using a product (laptop) "expert reviewers", and those who are beginners at using the product (laptop) as "beginner reviewers". Furthermore, we designate users who read reviews and who are experts for products (laptop) as " expert users", and users who read reviews and who are beginners for products (laptop) as "beginner users". Expert reviewers have details about features of the products and about their good points and bad points. Nevertheless, such detailed information is not good for beginner users. By contrast, reviews written by beginner reviewers have usability from the perspective of beginner users. This information is not good for expert users; however, beginner users regard the review as a high-value review. In this way, in the case of such a laptop computer, high-value reviews depend on the user type. We can understand hypothesis 2 as "False". However, almost all

Table 4. Rate of both positive and negative clauses

Product	The nNumber of users' evaluations				
	0	3 or more	5 or more	7 or more	10 or more
Toilet paper	0.291 (62/213)	0.346 (36/104)	0.407 (22/54)	0.429 (12/28)	0.667 (6/9)
Insect repellant	0.254 (54/213)	0.439 (58/132)	0.478 (33/69)	0.488 (20/41)	0.667 (12/18)
Laptop	0.300 (15/61)	0.455 (136/299)	0.486 (105/216)	0.488 (78/160)	0.460 (57/124)
Smartphone	0.385 (50/130)	0.521 (135/259)	0.561 (97/173)	0.594 (79/133)	0.679 (57/84)
Makeup base	0.477 (62/130)	0.528 (95/180)	0.553 (63/114)	0.625 (45/72)	0.654 (34/52)

Table 5. Rate of positive clauses

Product	Number of users' evaluations				
	0	3 or more	5 or more	7 or more	10 or more
Toilet paper	0.408	0.323	0.309	0.223	0.252
Insect repellant	0.368	0.245	0.199	0.198	0.195
Laptop	0.291	0.219	0.176	0.148	0.139
Smartphone	0.341	0.244	0.230	0.212	0.169
Makeup base	0.443	0.329	0.295	0.268	0.245

products show similar trends of high-value reviews. For products, such as laptop computers, for which there are expert users and novice users, the trends of high-value reviews differ.

Hypothesis 4. Reviews by expert reviewers have many neutral clauses.
After analyzing hypothesis 3, we were able to find that types of high-value reviews for some products depend on the reviewer type and the user type. Therefore, we propose hypothesis 4. We present an analysis of the relation as explained below.

1. We divide the reviews written about laptops into those of reviewers of three types. The three reviewer types are expert reviewers, general reviewers, and beginner reviewers. The expert reviewers know detailed information about laptop computers. The general reviewers know general information about laptop computers. Their knowledge about laptops is between expert and beginner. The beginner reviewers know almost no information about laptops. We determined the reviewer types using crowdsourcing. Of the crowd, 10 workers tagged reviewer types for reviews. They read a review and judged the reviewer type. We regard a reviewer type when judged by more than 6 out of 10 workers.

2. We determine the sentiment of clauses included in a review using crowdsourcing. The method of determining sentiment is the same as the method of determining reviewers.

3. We analyze the relation between reviewer types and high-value reviews.

Table 6. Rate of negative clauses

Product	Number of users' evaluations				
	0	3 or more	5 or more	7 or more	10 or more
Toilet paper	0.107	0.177	0.171	0.155	0.174
Insect repellant	0.091	0.151	0.189	0.185	0.126
Laptop	0.155	0.221	0.268	0.287	0.305
Smartphone	0.198	0.204	0.215	0.218	0.246
Makeup base	0.146	0.227	0.245	0.270	0.311

Table 7. Rate of neutral clauses

Product	Number of users' evaluations				
	0	3 or more	5 or more	7 or more	10 or more
Toilet paper	0.366 (78/213)	0.394 (41/104)	0.519 (28/54)	0.571 (16/28)	0.556 (5/9)
Insect repellant	0.399 (85/213)	0.659 (87/132)	0.681 (47/69)	0.659 (27/41)	0.722 (13/18)
Laptop	0.377 (23/61)	0.595 (178/299)	0.615 (133/216)	0.650 (104/160)	0.645 (80/124)
Smartphone	0.415 (54/130)	0.610 (158/259)	0.630 (109/173)	0.669 (89/133)	0.714 (60/84)
Makeup base	0.277 (36/130)	0.383 (69/180)	0.404 (46/114)	0.431 (31/72)	0.423 (22/52)

We regard both general reviewers and beginner reviewers as unfamiliar with laptops. Table 8 shows the rate of the reviewer types and the number of high-value reviews in each user's evaluation. Table 9 and Figure 6 show the average of neutral sentiment by reviewer type. Figure 7 and Figure 8 respectively show averages of positive/negative sentiment by reviewer type. In Figure 6, the difference in the ratio of neutral clauses between expert reviewers and non-expert reviewers might be readily apparent. Furthermore, comparing Figure 6 with Figure 7 and Figure 8, the ratio of positive and negative is almost identical to that of expert reviewers and non-expert reviewers. Then, we can verify hypothesis 4 as "True".

Table 8. Rate of reviewer types for laptop computers

Type of knowledge	reviewer type	Number of users' evaluations				
		0	3-4	5-6	7-9	10 and more
Deep knowledge	Expert	0.105 (4/38)	0.238 (65/273)	0.262 (51/195)	0.273 (39/143)	0.248 (27/109)
Not deep knowledge	General	0.684 (26/38)	0.553 (151/273)	0.528 (103/195)	0.510 (73/143)	0.486 (53/109)
	Beginner	0.211 (8/38)	0.209 (57/273)	0.210 (41/195)	0.217 (31/143)	0.266 (29/109)

Summary discussion

From these investigations, we can infer characteristics of high-value reviews as follows:

- High-value reviews have multiple sentiment clauses.

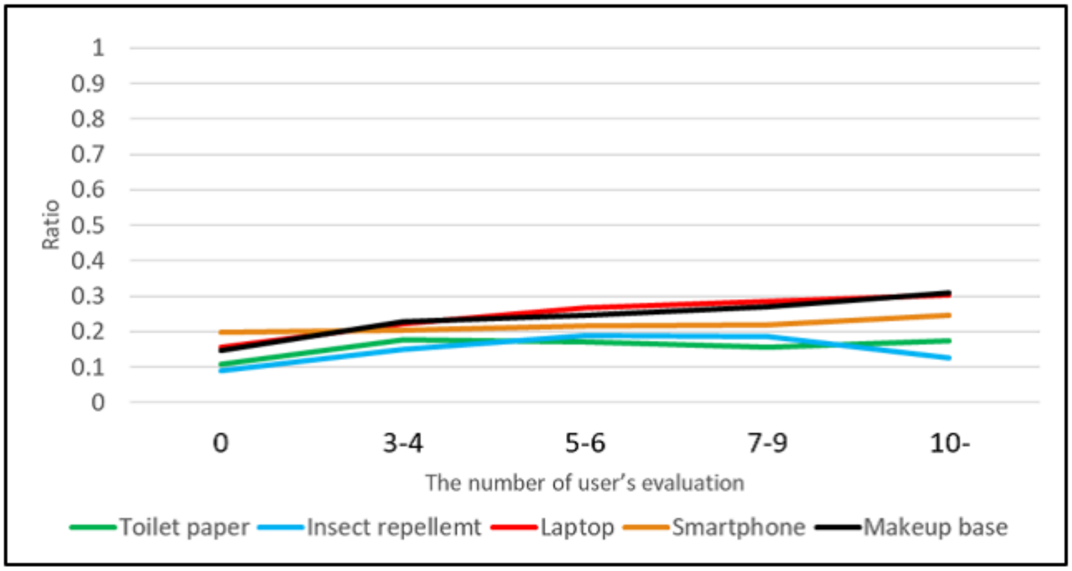

Fig. 4. Rate of both negative clauses.

Table 9. Average of neutral clauses by knowledge level for laptops

Type of knowledge	The number of users' evaluation				
	0	3-4	5-6	7-9	10 or more
Familiar with laptop	14.25	10.631	11.862	14.333	16.889
Not familiar with laptop	4.500	7.493	9.044	9.663	10.512

- Higher review ratings are associated with fewer positive clauses and more negative clauses.
- According to the type of product, high-value reviews depend on reviewer knowledge about the product.

4. Clause-level sentiment classification using CRF

As described in this Section, we measure a CRF to ascertain whether it is suitable for sentiment classification for clause-level reviews by comparing the baseline results to those of CRF. The baseline is dictionary-based sentiment classification. When we compare the methods, we also consider data granularity.

4.1. *Data*

When classifying the reviews, we use clauses in reviews that classify sentiment by our questionnaire as training data and evaluation data. We use product data of two types: daily necessities and home appliances from Amazon.[b] For differences in topics of review granularity, we particularly examine class-level topics and instance-level topics of reviews. We regard

[b] Amazon https://www.amazon.com

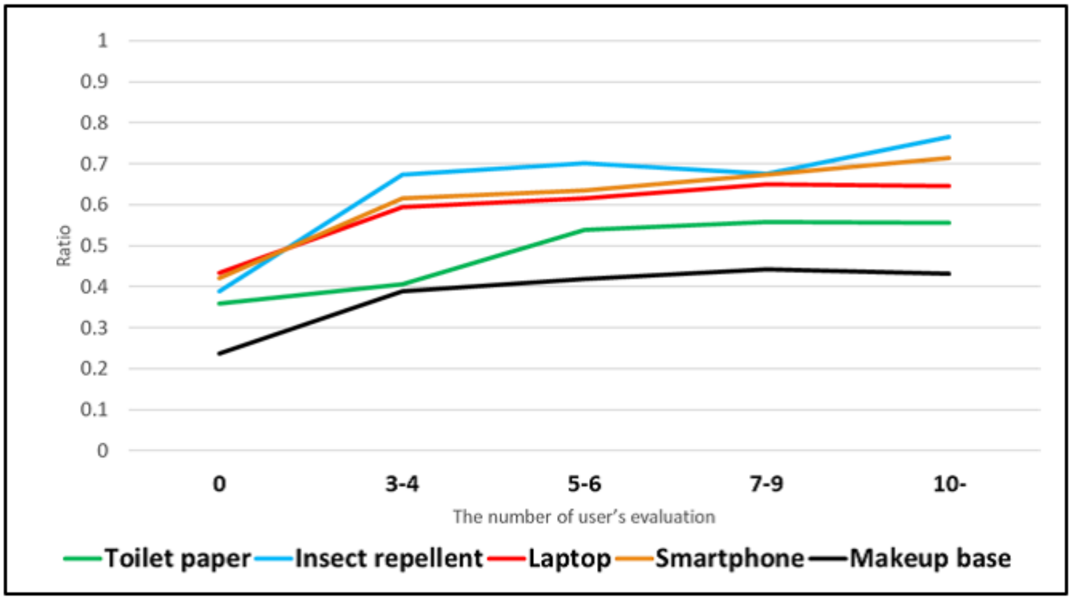

Fig. 5. Rate of neutral clauses.

Table 10. Number of class-level topic data

Data Name	Positive	Negative	Neutral	Kappa coefficient
Amazon_daily necessities	2966 (27%)	1512 (14%)	6705 (60%)	0.479
Amazon_home appliances	4094 (22%)	2995 (16%)	11320 (61%)	0.485
Fuman (Complaining data)	69 (2%)	2257 (50%)	2152 (48%)	0.435

reviews of daily necessities and home appliances as class-level topics of reviews. We regard reviews of toilet paper, detergents, and insecticides as instance-level topics of the daily necessities class. We also analyze reviews of laptop PCs, robot vacuum cleaners, and smartphones as instance-level topics of the home electronics class. Tables 10 and 11 show the number of data: the number of clauses and each kappa coefficient is a statistic that is inter-rater agreement for qualitative items.

This study was conducted to classify positive/negative/neutral sentiment automatically from reviews. However, a general review such as an Amazon review includes less negative opinion information as in Tables 10 and 11. We examine data that include large amounts of negative opinion data. Such data benefit our purposes. We analyze not only general review data but also large amounts of negative opinion data. Then we analyze "Amazon" data as general review data and "Fuman [23]" data as large amounts of negative opinion data. The Fuman data are submitted by users along with some complaint. Among them are formatted texts in which users complain about widely various topics such as products and services, society, and even everyday life such as "snoring is annoying."

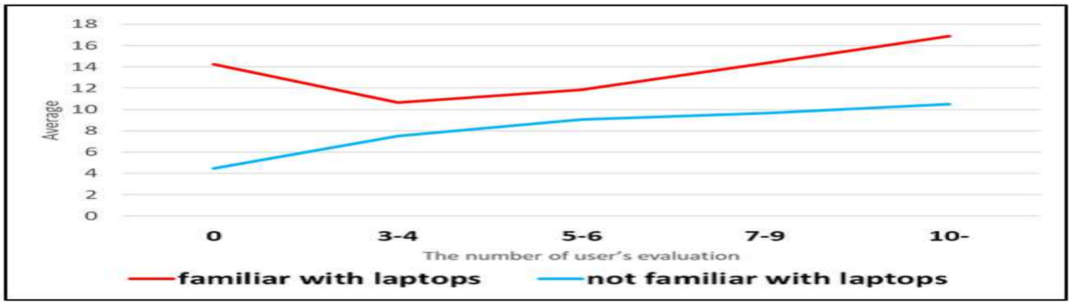

Fig. 6. Average of neutral clauses for laptops based on the user type.

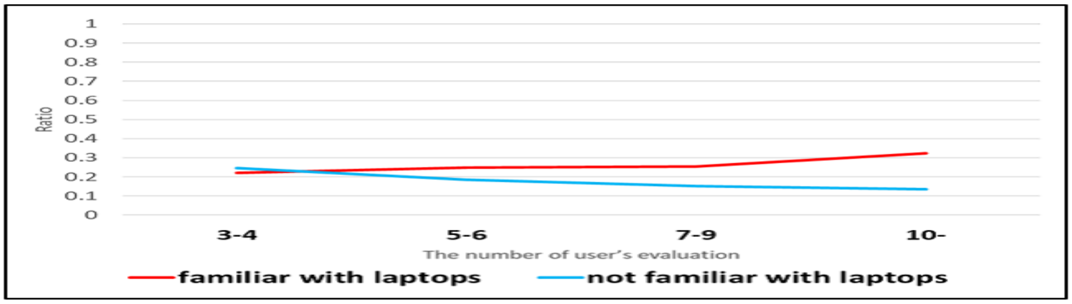

Fig. 7. Average of positive clauses for laptops based on the user type.

Furthermore, we used crowdsourcing for questionnaires to find correct answers. We conducted sentiment classification by 10 workers against 33,587 clauses divided by a Japanese Dependency and Case Structure Analyzer KNP.

First, the workers read each clause and evaluated whether the clause applies as positive, negative, or neutral. We regard the correct sentiment of a clause as one that is judged by 6 or more people out of 10 indicated workers. When no evaluation on the sentiment axis reaches a majority, the sentiment of the corresponding clause is regarded as neutral because it has no definite sentiment. We calculate the kappa coefficient shown in Tables 10 and 11. For the class-level topic experiment, we use 1200 reviews in each class-level topic Amazon review as training data and 300 reviews in each class-level topic Amazon review as test data. Furthermore, in the instance-level topic experiment, we use 400 reviews in each instance-level topic Amazon review as training data and 100 reviews in each instance-level topic Amazon review as test data.

4.2. *Baseline*

When measuring the benefits of the proposed clause-level sentiment classification using CRF, we compare CRF with the baseline. The baseline is a dictionary-based sentiment classification.

Dictionary

We use the word-sentiment polarity correspondence table created by Takamura et al. [24], which includes 55,125 words: nouns, verbs, adjectives, auxiliary verbs, and adverbs. Each

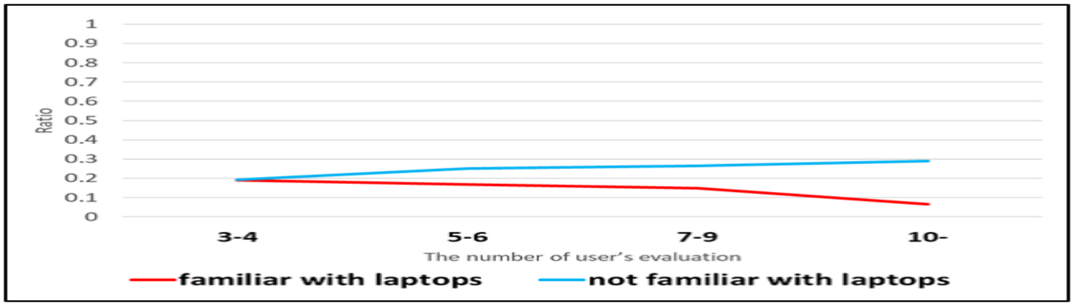

Fig. 8. Average of negative clauses for laptop based on the user type.

Table 11. Number of analysing instance-level topic data

Data	Positive	Negative	Neutral	Kappa coefficient
Amazon_toilet paper	955 (31%)	442 (14%)	1721 (55%)	0.473
Amazon_detergent	1234 (28%)	700 (16%)	2492 (56%)	0.504
Amazon_insecticide	777 (24%)	370 (12%)	2492 (64%)	0.460
Amazon_laptop PC	1054 (21%)	858 (17%)	3187 (62%)	0.464
Amazon_robotic vacuum cleaner	1922 (25%)	1051 (14%)	4628 (61%)	0.479
Amazon_smartphone	1118 (20%)	1086 (19%)	3505 (61%)	0.513

word has a value from -1 to 1 as a sentiment polarity value. The value of -1 is the most negative value; 1 is the most positive value.

Calculating sentiment classification using the dictionary

The value of clause-level sentiment polarity PA is the average of the sentiment polarity values of each word included in the clause-level using the following expression. In that expression, k denotes the number of words included in the word sentiment polarity correspondence table in the clause. Also, P_i is the value of the polarity of word i, which is the i-th word in the clause. The value of words that are not included in the word sentiment polarity correspondence table becomes 0.

$$PA = \begin{cases} \frac{\sum_{i=1}^{k}(P_i)}{k} & (k >= 1) \\ 0 & (k = 0) \end{cases}$$

After we calculating the value of PA, we define the clause polarity, which is positive/negative/neutral depending on the value of PA as follows. Furthermore, the threshold of PA is that we set about three divisions of the range of -1 to 1, which is the polarity value of the dictionary

$$Clause\ polarity = \begin{cases} Positive & (PA > 0.3) \\ Negative & (PA <= -0.3) \\ Neutral & (-0.3 < PA <= 0.3) \end{cases}$$

4.3. *CRF*

We infer that conditional random fields (CRF)[25] are a benefit to calculate clause-level sentiment clustering because it is reasonable to infer not only the clause that is calculated from sentiment clustering but also the relation with the clause around it. Specifically, we use CRFsuite[26], which is a kind of linear-chain CRF.

For the training algorithm, we use Limited-memory BFGS (L-BFGS) method [27], which is suitable for solving convex optimization efficiently. Parameters for CRFsuite we used were the following: Coefficient parameter c1 was set to 1.0, which represents the coefficient for L1 regularization. Additionally coefficient parameter c2 was set as 0.001, which indicates the coefficient for L2 regularization. For training, we use not only features of a clause, but also features of two subsequent clauses and the prior clause.

4.4. *Features*

We use the original form of each word, part of speech, polarity value, and the number of polarity inversion words in the target clause as feature data. The polarity inversion word refers to a word that can influence the polarity, which is the sentiment of the later clause such as "But" or "However." For this study, we use 39 conjunctions and connective particles that have opposite meaning. They are in the thesaurus database, which is published by the National Institute of Japanese Language footnote https://www.ninjal.ac.jp/. When we calculate the sentiment polarity of a word, we use the word sentiment polarity correspondence table created by Takamura et al. [24]. Specifically, we determine x as the polarity value of a word according to n. Also, n is the value of word sentiment polarity correspondence table, as described below.

$$x = \begin{cases} 2 & (n > 0.5) \\ 1 & (0.5 > n > 0) \\ -1 & (0 > n > -0.5) \\ -2 & (n < -0.5) \end{cases}$$

4.5. *Results and Discussion*

The experiment was conducted to show that CRF is a benefit for sentiment classification of clause-level review based on comparison with our proposed dictionary-based method. Table 12 shows the results of our experiments, We conducted experiments of two types, with class-level topics and instance-level topics of reviews.

Instance-level topic data

Figure 9 and Figure 10 present results of instance-level experiments. Comparison of the results of precision, recall, and F-measure between CRF and the baseline for the instance-level topic data reveals that almost all results of CRF are higher than the results of the

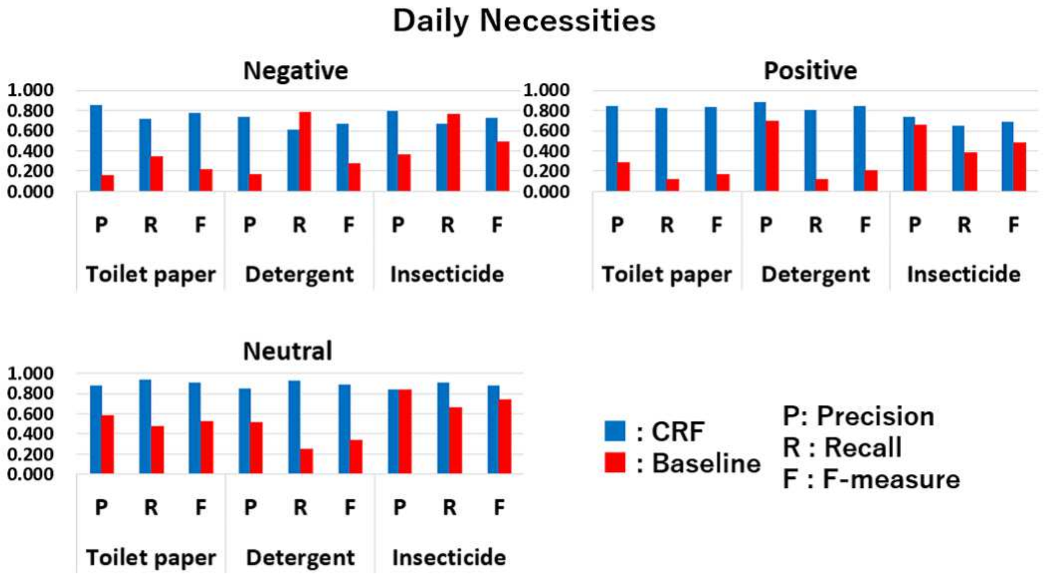

Fig. 9. Result of Instance-level topic of Daily Necessities Reviews

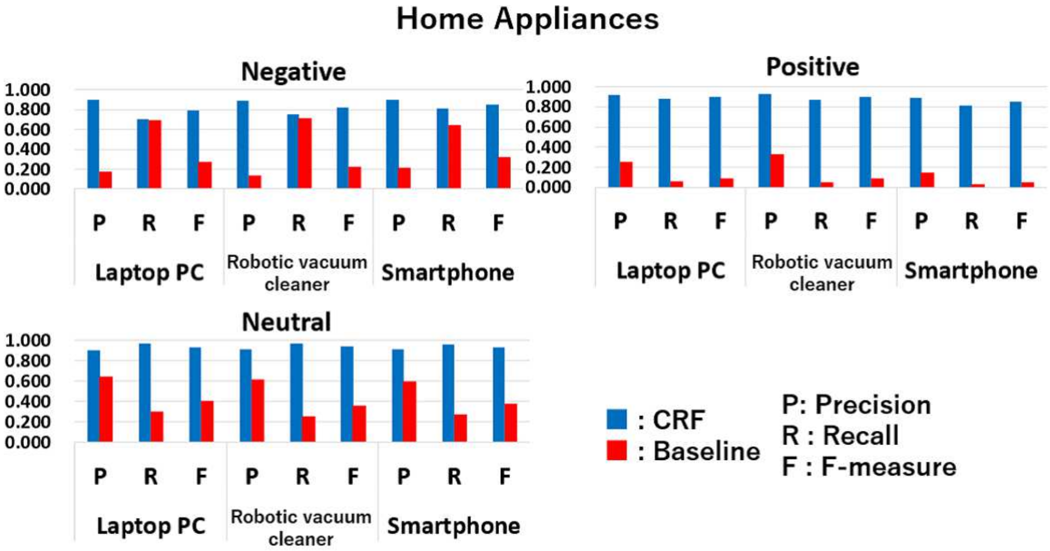

Fig. 10. Result of Instance-level topic of Home Appliances Reviews

baseline. The reason is that the baseline does not consider the context of the clause, although CRF considers the context before and after the clause. Table 13 presents an example of consideration of clause level sentiment. The baseline judges that clause 2 is neutral, but CRF judges that clause 2 is negative because it considers the clause context. In this way, we can ascertain that CRF is more suitable for clause-level sentiment classification than the baseline.

The recall of negative is lower than the precision in CRF analysis because the CRF can

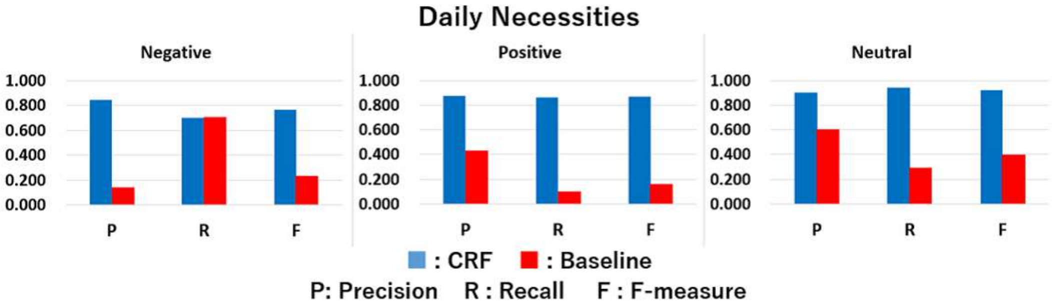

Fig. 11. Result of Class-level topic of Daily Necessities Reviews

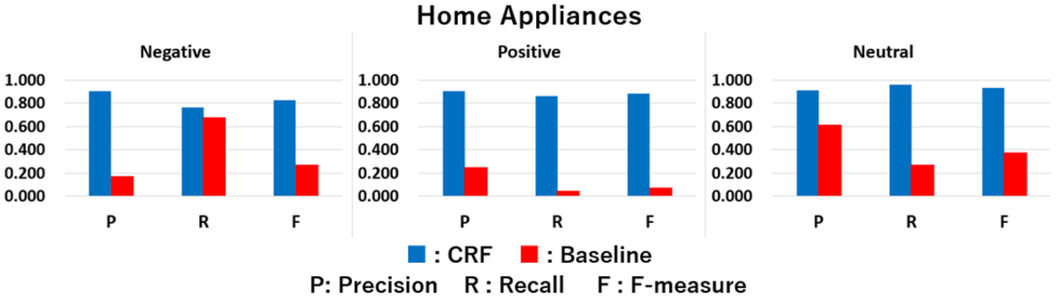

Fig. 12. Result of Class-level topic of Home Appliances Reviews

analyze direct negative expressions such as “The usability of the vacuum cleaner is very bad.” and “The cost performance is very bad.” Nevertheless, it is incapable of analyzing indirect negative expressions such as “The cleaner smells like fish.” and “The keyboards sounds clattery.”.

Class-level topic data

Figure 11 and Figure 12 presents results of class-level experiments. Results of class-level topic data and results of CRF are higher than the results of the baseline, which is true also for results of instance-level topic data. In addition, the recall of CRF analysis is lower than its precision for the same reason as that for instance-level topic data. The CRF can analyze direct negative expressions such as “Water absorption is bad and it is easy to break,” but it cannot analyze indirect negative expressions such as “I did not know the efficacy.”

Results presented above show some probability that, when using CRF, the sentiment classification of review by the clause-level is more suitable than the baseline because CRF incorporates consideration of the context of a clause. However, CRF cannot indirectly assess sentiment data. It is necessary to respond to such data in future studies.

5. Conclusion

As described in this paper, we undertook two investigations: (1) we assessed the relation between high-value reviews and their sentiment based on clauses and (2) we evaluated clause-level classification sentiment for reviews. For investigation (1), the high-value reviews were found to be independent of the star rating. Instead, they were found to depend on the user evaluation. Our analysis of sentiment was conducted on three axes: positive, negative, and

Table 12. Comparing CRF and Baseline

Data	Label	CRF			Baseline		
		Precision	Recall	F measure	Precision	Recall	F measure
Toilet paper	Positive	0.847(265/313)	0.823(265/322)	0.835	0.286(36/126)	0.112(36/322)	0.161
	Negative	0.851(212/249)	0.719(212/295)	0.779	0.151(101/609)	0.342(101/295)	0.210
	Neutral	0.884(893/1010)	0.935(893/955)	0.909	0.584(454/777)	0.475(454/955)	0.524
Detergent	Positive	0.883(212/240)	0.806(212/263)	0.843	0.689(31/45)	0.118(31/263)	0.201
	Negative	0.741(80/108)	0.611(80/131)	0.669	0.168(102/606)	0.779(102/131)	0.277
	Neutral	0.851(468/550)	0.929(468/504)	0.888	0.514(127/247)	0.252(127/504)	0.338
Insecticide	Positive	0.736(178/242)	0.645(178/276)	0.687	0.658(104/158)	0.377(104/276)	0.479
	Negative	0.794(196/247)	0.664(196/295)	0.723	0.360(226/627)	0.766(226/295)	0.490
	Neutral	0.843(913/1083)	0.912(913/1001)	0.876	0.840(661/787)	0.660(661/1001)	0.739
Laptop PC	Positive	0.919(205/223)	0.884(205/232)	0.901	0.250(12/48)	0.052(12/232)	0.086
	Negative	0.903(130/144)	0.703(130/185)	0.790	0.169(128/758)	0.692(128/185)	0.271
	Neutral	0.905(707/781)	0.967(707/731)	0935	0.640(219/342)	0.300(219/731)	0.408
Robotic-vacuum-cleaner	Positive	0.933(334/358)	0.872(334/383)	0.901	0.328(19/58)	0.050(19/383)	0.086
	Negative	0.888(151/170)	0.759(151/199)	0.818	0.129(142/1103)	0.714(142/199)	0.218
	Neutral	0.916(936/1022)	0.967(936/968)	0.941	0.617(240/389)	0.248(240/968)	0.354
Smartphone	Positive	0.893(191/214)	0.813(191/235)	0.851	0.143(6/42)	0.026(6/235)	0.043
	Negative	0.902(294/326)	0.814(294/361)	0.856	0.214(231/1077)	0.640(231/361)	0.321
	Neutral	0.909(938/1032)	0.961(938/976)	0.934	0.592(268/453)	0.275(268/976)	0.375

Table 13. Example: Consideration of context

Clause No.	Text	Correct data	Prediction result	
			CRF	Baseline
1	I expected that I can clean the dirt with only the detergent	Positive	Positive	Positive
2	but I have to scrub with the sponge	Negative	Negative	Neutral
3	to clean the dirt	Neutral	Neutral	Neutral

neutral. When we investigate the relation between high-value reviews and their sentiment, we generate four hypotheses and discuss the results individually. From the results, we can infer that the characteristics of high-value reviews are such that (a) high-value reviews have multiple sentiment clauses, (b) higher review ratings are associated with less positive clauses and the more negative clause, and (c) by the type of product, high-value reviews depend on reviewers' knowledge of the product. In investigation (2), we compared CRF with a dictionary-based baseline. Results show that the CRF is beneficial for clause-level sentiment classification. We conducted an experiment to compare the proposed system with the baseline. From the experiment results, we can infer that our proposed method is beneficial for extraction of multiple sentiments from a review.

In the near future, the following must be considered:

- Calculating sentiment based on the order of sentiment clauses.
- Analyzing relations between high-value reviews and the types of customers who read reviews.
- Investigating other products.
- Developing a review recommendation system based on research results.

Acknowledgements

This work was partially supported by Research Institute of Konan University, and by JSPS KAKENHI Great Number 19H04221, 19H04218 and 20K12085.

References

1. Ministry of Internal Affairs and Communications (2016), *Chapter1 Section4 Multifaceted ICT Contribution to the Economy and Society*, the 2016 WHITE PAPER on Information and Communications in Japsn. pp.17-19.
2. M. Ando, S. Sekine (2014), *What is written in the review? What does a reader read in the review?*, The proceedings of the annual meeting 2014 of the association for natural language processing, pp. 884-887(in Japanese).
3. Wu, Derek., Wang, Hongning (2017), *ReviewMiner: An Aspect-based Review Analytics System*, The proceedings of SIGIR2017, pp.1285-1288.
4. Patrali, Chatterjee (2001), *Online Reviews: Do Consumers Use Them?*, Advance in Consumer Research Vol. 28, pp. 129-134.
5. Z. Nanli, Z. Ping, L. Weiguo and C. Meng (2012), *Sentiment analysis: A literature review*, The proceedings of International Symposium on Management of Technology (ISMOT2012), pp. 572-576.
6. Yohan Jo , Alice H. Oh (2011), *Aspect and sentiment unification model for online review analysis*, The proceedings of the fourth ACM international conference on Web search and data mining(WSDM 2011), pp. 815-824.
7. Devika M D, Sunitha C, Amal Genesh (2016), *Sentiment Analysis: A Comparative Study On Different Approaches*, The proceedings of the Fourth International Conference on Recent Trends in Computer Science & Engineering (ICRTCSE 2016), pp. 44-49.
8. Duy Khang Ly, Kazunari Sugiyama, Ziheng Lin and Min-Yen Kan (2011), *Product review summarization from a deeper perspective*, The proceedings of the 11th annual international ACM/IEE Joint Conference on Digital Library(JCDL2011), pp. 311-314.
9. C. Liu, W. Hsaio, C. Lee, G. Lu and E. Jou (2012), *Movie Rating and Review Summarization in Mobile Environment*, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 3, pp. 397-407.
10. Li Zhuang, Feng Jing, Xiao-Yan Zhu (2006), *Movie review mining and summarization*, The proceedings the 15th ACM international conference on Information and knowledge management(CIKM 2006), pp. 43-50.
11. Abulaish M.uhammad, Jahiruddin, Doja Mohammad Nalumd, Ahmad Tanvir (2009), *Feature and Opinion Mining for Customer Review Summarization*, The proceedings of the third international conference on Pattern Recognition and Machine Intelligence(PReMI 2009), pp. 219-224.
12. Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang and Ming Zhou (2007), *Low-Quality Product Review Detection in Opinion Summarization*, The proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 334-342.
13. Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang and Hao Yu (2010), *Structure-aware review mining and summarization*, The proceedings of the 23rd International Conference on Computational Linguistics(COLING '10), pp. 653-66.
14. Sudhof, M., Emilsson, A.G., Maas, A.L., Potts, C. (2014), *Sentiment expression conditioned by affective transitions and social forces*, The proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp.1136–1145.
15. Zhang, K., Xie, Y., Cheng, Y., Honbo, D., Downey, D., Agrawal, A., Liao, W., Choudhary, A. (2012), *Sentiment identification by incorporating syntax, semantics and context information*, The proceedings of SIGIR 2012, pp.1143–1144.

16. Zhao, J., Liu, K., Wang, G. (2008), *Adding redundant features for CRFs-based sentence sentiment classification*,The proceedings of Conference on Empirical Methods in Natural Language Processing(EMNLP), pp.117–126.
17. Patra, B.G., Mandal, S., Das, D., Bandyopadhyay, S. (2014), *JU_CSE: A Conditional Random Field (CRF) Based Approach to Aspect Based Sentiment Analysis*, The proceedings of International Workshop on Semantic Evaluation(SemEval), pp370-374.
18. Yang, B., Cardie, C. (2014), *Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization*, The proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics(ACL), pp.325–335.
19. Choi, Y., Cardie, C. (2008) *Learning with compositional semantics as structural inference for subsentential sentiment analysis*, The proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.793–801.
20. Paltoglou, G., Thelwall, M. (2013), *More than Bag-of-Words: Sentence-based Document Representation for Sentiment Analysis*, The proceedings of Recent Advances in Natural Language Processing(RANLP), pp.546–552.
21. Socher, R., Perelygin, A, Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts ,C. (2013), R̂ecursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, The proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1631–1642.
22. Kurohashi, S., Nagao, M. (1994), *KN Parser : Japanese Dependency/Case Structure Analyzer*', The proceedings of The International Workshop on Sharable Natural Language Resources, pp.48-55.
23. Mitsuzawa, K., Tauchi, M, Domoulin, M., Nakashima, M., Mizumoto, T. (2016), *FKC Corpus: a Japanese Corpus from New Opinion Survey Service*, The proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results, pp.11-18.
24. Takamura, H., Inui, T., Okumura, M. (2005), *Extracting Semantic Orientations of Words using Spin Model*, The proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005) , pp. 133–140.
25. Lafferty, J., McCallum, A., Pereira, F.C.N. (2001), *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, The proceedings of International Conference on Machine Learning(ICML), pp.282–289.
26. Okazaki, N. (2007), *CRFsuite: A fast implementation of Conditional Random Fields*, http://www.chokkan.org/software/crfsuite/, Accessed 11 May 2018.
27. Jorge Nocedal (1980), *Updating Quasi-Newton Matrices with Limited Storage*, Mathematics of Computation. 35. 151, pp.773–782.