What Makes a Review Encouraging: Feature Analysis of User Access Logs in a Large-scale Online Movie Review Site

Kakeru Ito*
Tokyo Institute of Technology
Tokyo, Japan
kakeru@sw.it.aoyama.ac.jp

Yoshiyuki Shoji
Aoyama Gakuin University
Kanagawa, Japan
shoji@it.aoyama.ac.jp

Sumio Fujita
Yahoo JAPAN Corporation
Tokyo, Japan
sufujita@yahoo-corp.jp

Martin J. Dürst
Aoyama Gakuin University
Kanagawa, Japan
duerst@it.aoyama.ac.jp

ABSTRACT

This paper reveals the characteristics of the reviews that encourage readers to watch the reviewed movie by analyzing large-scale access log data. We assume that some of the reviews that users saw just before they clicked the links to a streaming site contain factors that help users decide whether they watch that movie. Our method used a random forest classifier trained to determine whether a review encouraged a movie-watching behavior. We conducted feature importance-based analysis using three types of features: review itself, item, and reviewer. We analyzed 70,000 user behaviors from Yahoo! Movies (a movie review site in Japan) and Gyao! (a movie streaming site in Japan). Through a cross-validation experiment, the classifier was able to classify encouraging reviews with an F-score of 0.78, and mainly the features about the item contributed to the classification performance. An additional subjects experiment confirmed that these features contribute to the review's usefulness.

KEYWORDS

Online Review, Access Log Analysis, Random Forest, Feature Study

1 INTRODUCTION

The total number of movies released by 2021 exceededs 570,000 [1]. Assuming each movie is roughly two hours long, it would take 130 years to see all movies. No one can watch all these movies in his or her lifetime. People need to continuously make quick decisions about which movies to watch and which ones not to watch while they are alive.

Online review sites are an essential information resource to help us make such decisions in our daily life. We regularly read reviews about products on shopping sites, about videos on streaming sites, about facilities on maps, about schools, doctors, individuals, and so on. On the other hand, not all of the reviews posted on these review sites are useful for decision-making. Various people write reviews; some are well-written, some are useless. It is difficult to pick out only the information that can be used efficiently for decision-making from review sites.

Nowadays, there is a large number of reviews for each product. It is clear that reading many reviews requires much time to make a decision. We consider estimating the helpfulness for decision-making of a review by analyzing how the review encouraged users' page transition. Many studies discussed the usefulness of reviews and what characteristics make a review useful. However, in general, these studies derive the degree of usefulness of a review from whether readers clicked the "helpful" button [6, 18, 22]. The usefulness of these reader polls does not necessarily equate to their ability to be used in actual decision-making. Consider when readers will press the "helpful" or "unhelpful" button. For example, some paid trolls may press the "helpful" button on a highly rated review of a product they support as an act of stealth marketing. This kind of voting behavior is not necessarily valuable for decision-making from the average user's perspective. Therefore, we suggest that users not click the "helpful" button when they read a really useful review. In this study, we assume that users who read really useful reviews would purchase the product before clicking the "helpful" button.

In this study, we analyzed actual large-scale log data in order to estimate the degree that reviews are useful for decision-making. First, we collected access logs of people actually watching movies on the streaming site immediately after reading some reviews. The actual access log was taken from Yahoo! Movies[2], one of the biggest online movie communities in Japan, and Gyao[3], a well-known movie streaming service in Japan, were used for the analysis (see Figure 1). Next, we learned a random forest classifier that separates the reviews that encouraged movie-watching behavior from other reviews. Three different types of features represented a review;

- Features related to the **review itself** are characteristics contained in an individual review, such as the review's rating, the style of writing, and the vocabulary used,
- Features related to the reviewed **item** (*i.e.*, movie) that is the subject of the review consists of metadata such as director, movie category, and reputation such as the number of reviews and the overall movie rating, and
- **Reviewer** related features are features of the reviewer who wrote the review. They consist of information about the reviewer's experience, such as the number of reviews written so far, and for how long they have been writing reviews, or their grading bias.

*Kakeru Ito contributed to this research while at Aoyama Gakuin University until March 2020

[1] IMDb Statistics - Press Room - IMDb: https://www.imdb.com/pressroom/stats/

[2] Yahoo! Movies: https://movies.yahoo.co.jp/

[3] Gyao! https://gyao.yahoo.co.jp/

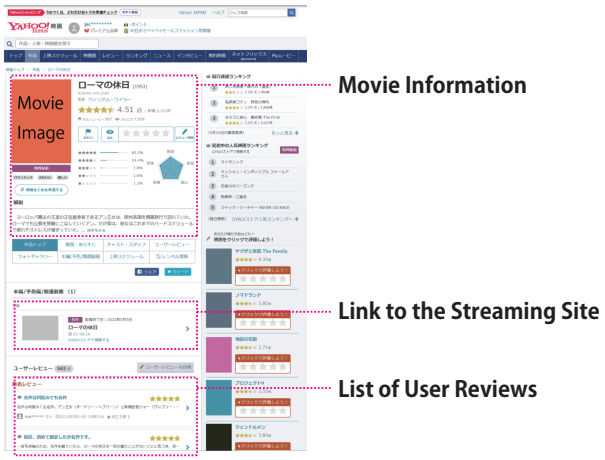

Figure 1: Screen shot of a movie detail page in Yahoo! Movies in Japan (Movie images were masked due to the Copyright). Users can move to the streaming site after reading reviews

In a random forest classifier, the contribution of each feature can be calculated as the importance when performing classification. Features with high importance are expected to be a substantial factor in determining whether the review encourages people's movie-watching behavior. Finally, we conducted a subject experiment with a questionnaire about reviews with these characteristics. Subjects answered how much reading a review made them want to see the reviewed movie.

There are two practical advantages to using such a classifier for analysis. First, it can extract reviews that are actually encouraging. Since the confidence of the classification results can rank the extracted reviews, the review site will be able to show only the reviews that are likely to be effective for decision-making. Second, features with different ranges of values and forms of expression can be compared integrally with their contribution to the classification. Some review-related features take real values, such as the number of characters, while others take discrete values, such as the number of stars. These features cannot be directly compared or analyzed.

The structure of this paper is shown below. This section described the background and purpose of our research. In section 2, related research fields with similar objectives and specification techniques are introduced and discussed. Section 3 describes the analysis method proposed in this research. Section 4 shows the method and result of the actual analysis using real access logs of Yahoo! Japan and Gyao. Section 5 describes the subject experiment and its results. In section 6, we discuss the results, and in section 7, we conclude and discuss future prospects.

2 RELATED WORK

This study identifies reviews that induce consumption behavior by using various features. This section introduces and discusses related studies that target reputational information or eWOM (electric Word of Mouse). We also introduce the features associated with reviews (*i.e.*, review targets, reviewers) and discuss the differences and positioning of this research.

2.1 Review's Usefulness

Many studies have been conducted to determine the relationship between the characteristics of a review and its usefulness.

Zhou *et al.* [22] define the usefulness of a review as the degree to which reading the review improves one's ability to evaluate the product. They categorize the features of reviews into two types: numerical features and textual features. Furthermore, they discuss the hypothesis that the two types of features influence each other. As numerical features, they used metadata that the reader can obtain before reading the text, such as the length of the text and the rating. As textual features, they used information that can only be obtained by reading the text, such as the text sentiment or the average number of words in a sentence. Experiments showed that longer reviews give readers more confidence and are more likely to be deemed useful.

Among the numerical features of reviews, there have been many studies on the relationship between rating and usefulness. However, conflicting conclusions have been reported. Pan *et al.* [18] stated that reviews with higher scores are more likely to be perceived as useful reviews. On the contrary, Chua *et al.* [6] stated that reviews with higher ratings are less likely to be perceived as useful reviews.

In contrast to these two results, Mudambi *et al.* [16] showed that there is a non-linear relationship between the usefulness of a review and its rating. They reported that reviews with high and low ratings are less likely to be perceived as useful, and reviews with average ratings are most likely to be judged as useful.

Other textual features, such as the polarity of the text sentiment, have also been studied in terms of their effects on consumer purchasing behavior [2, 5]. The inconsistent effect of text polarity on the usefulness of a review has also been reported by Hao *et al.* [9] .

Hong *et al.* [10], in a meta-analysis of studies on the characteristics that affect the usefulness of reviews [12, 19], showed that the complexity of the review text and the time elapsed since the review was posted might have a significant effect on the usefulness of the review.

In a similar study on the usefulness of reviews, Chen *et al.* [3] pointed out that reviews that are rated as "Helpful" by other users have a more decisive influence on readers' purchase intentions. However, it is rare for ordinary users to click the "Helpful" button. Therefore, it is also pointed out that the indicator of evaluation from users is insufficient for judging usefulness [13, 21]. It has also been pointed out that there is a bias based on the timing of the posting of reviews. A newly posted review does not have enough time to be evaluated by other users.

2.2 Effect of Review Target

In this study, the characteristics of the items to be reviewed are also used for classification. For review sites, it is known that the characteristics of the items affect the behavior of users.

In a survey of eWOM, Nelson *et al.* [17] proposed a method to analyze two types of products, search goods and experience goods. Following this analysis, Mudambi *et al.* [16] mention that the type of product affected the usefulness perceived by review readers. Huang

et al. [11] found a phenomenon that sentences including subjective impression affect the reader's usefulness judgment, and objective reviews are judged as useless in the case of reading reviews of experience goods. Luan *et al.* [14] conducted an experiment using an eye-tracking device and compared the results with those of an experiment using a questionnaire to verify the results presented above. The results of gaze analysis showed that users' browsing behavior varies not only by product type but also by review content.

In this study, we analyzed movie reviews. It is difficult to judge the content of a movie by its specifications and performance, and it is impossible to evaluate its quality without experiencing it. Therefore, it is close to an experience good, and as in these analyses, we include subjective impressions and other characteristics. Also, as with eye gaze, the analysis in this study is based on the actual behavioral logs of users, not polls or questionnaires.

2.3 Reviewer's Authority

In addition to the reviews themselves, the impact of the reviewer on the reviews' usefulness has been discussed. Connors *et al.* [7] argued that when a reader judges the usefulness of reviews, a review that suggests it is written by an expert tends to be judged as more useful.

Studies have also been conducted on the relationship between a review's usefulness and the extent to which reviewers disclose their profiles. The degree of disclosure and usefulness are known to be positively correlated [1, 8]. As an example, Cheung *et al.* [4] analyzed what kind of reviewers are trusted through a large-scale survey. Depending on the data set used in the study, the amount of profile information a reviewer can disclose varies.

Different services have different forms that can be filled out for profiles, and in some cases, there are completely anonymous review sites. In our study, which uses the Yahoo! Movies dataset, reviewers cannot disclose much profile information. The readers can only see review history, screen names, and such systematic information. Thus, different review sites may have different opportunities and tendencies to judge the authority of reviewers.

3 ANALYSIS METHOD

Our method trained a random forest classifier that estimates whether a certain review induces a person to take action. This analysis method can be used for general review sites, but we will explain the method using movies as an example since we actually analyzed movie reviews.

Three kinds of features can roughly characterize a certain review; the information contained in a specific review, information about the author who wrote the review, and information about the movie that is the target of the review. The proposed method transforms a given review into a feature vector of 149 dimensions. Then, the number of people who actually watched the movie right after reading the review was aggregated and assigned as the correct answer label. Nowadays, most review sites include a link from the movie review page to the viewing page of an affiliated online streaming site. The method measures the click-through rate of this link, and extracts the reviews where an elevated ratio of readers moved to the streaming site. We used them as encouraging reviews, *i.e.*, as positive examples for the training. Reviews in this dataset were classified by random forest classifier, and the contribution of each feature or set of features is analyzed.

Table 1: List of features in the review feature vector

Type	Hyposesis	Feature	
Review itself	Amount of information	Text length	1
		Title length	1
		Spoiler tag	1
	Readability	# of words per sentence	1
		Frequency of line breaks	1
	Review content	Impression tags	20
		Topic	50
		Overall grading	1
		Rating by viewpoint	5
	Popurality	View count	1
		# helpful	1
Item	Metadata	Running time	1
		Release date	1
		Title length	1
	Item content	Category	15
		# of director's films	1
		Rating (age reccomendation)	1
	Popularity	Total # of reviews	1
		# fav	1
		# viewed	1
	Reputation	Total rating	1
Reviewer	Enthusiasm	# review posted	1
		Total # of characters	1
		# of movies watched	1
		# of helpful clicked	1
		Average # of characters	1
	Informativeness	% of reviews with spoiler tags	1
		Average view count	1
	Authority	Grading variance by viewpoint	1
		Average of tags attached	1
		# helpful received	1
Total # dimension			149

3.1 Features of Review Itself

The first type of feature concerns the content of the review itself. For example, the writing quality of the review itself is one of the essential indicators of whether the review is useful for decision-making. Therefore, we characterize the reviews by their amount of information, readability, content, and popularity. We hypothesize about each of these factors and represent them as 115-dimensional feature values.

We focused on the amount of information contained in the review. We used the number of characters in the review, the number of characters in its title, and the spoiler flag as features focusing on the information content of the review. The longer the review, the more content it has. It is more likely to contain helpful descriptions. The same goes for reviews with long review titles. The spoiler flag is 0 or 1, indicating whether the review contains the story's core. Reviewers can arbitrarily set this flag when they submit a review. The website does not display the text for posts containing spoilers until the viewer clicks on it.

We considered the readability of the review; an easy-to-read review must help readers make their decision [22]. Therefore, we

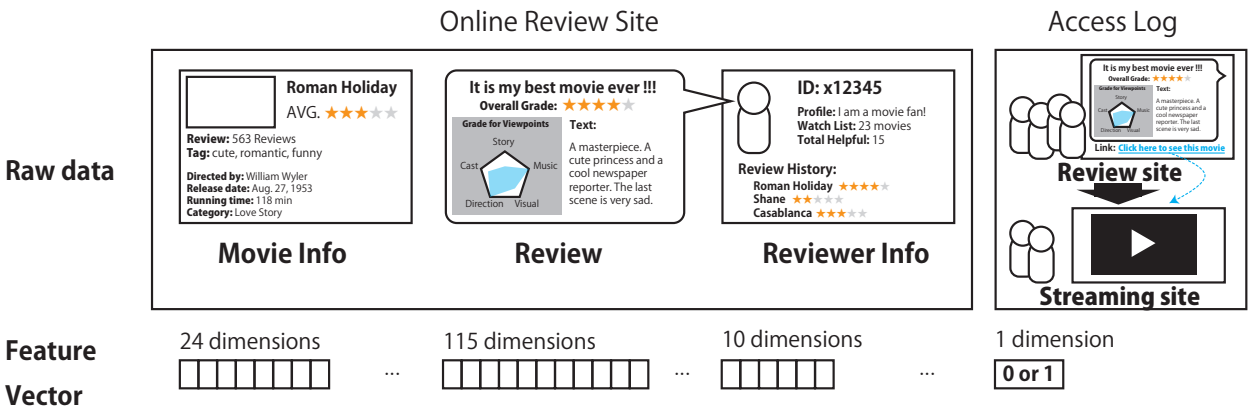

Figure 2: Outline of our analysis method. We created dataset consist of 149 dimensional feature vectors of the reviews, and the answer label that represent the encouragingness of a review.

used the length of a sentence (*i.e.*, the average number of words in a sentence, separated by periods or exclamation marks), and the number of linebreaks as readability indicators.

To represent the content of the review, we used four features: The impression tag, the topic of the words in the review text, the overall rating, and the rating by individual perspectives. The impression tag is a function that allows reviewers to add tags when they post a review. Specifically, reviewers can select zero or more tags from 20 types. Tags contain both impressions and movie categories, such as tearful, fantasy, and cool. Impression tags were used as binary features of 20 dimensions.

To indicate what words are used in the text of the review, we used LDA (Latent Dirichlet Allocation) to transform the words in the text into topics consisting of 50 dimensions. LDA is a method based on topic models and is commonly used for dimensionally compression of document vectors. As a training corpus for LDA, we used all review texts from our dataset.

Essentially, in LDA, a single document belongs probabilistically to multiple topics. This set of features accounted for more than one-third of the entire feature vector size in this experiment. Through our preliminary experiments, we found that when each of the 50 dimensions took real values, the classifier tended to be affected by the specific dimension in the topic, and the accuracy decreased. For this reason, we assumed that a review sentence belongs to a single topic and transformed the review content into a one-hot vector.

The rating of a review is also an important characteristic of the content of the review. In this review site, the rating is represented as the number of stars (*i.e.*, from one star to five stars). A review with five stars and a review with three stars are considered to have different degrees of usability for decision making. Some of the review sites allow reviewers to give a movie a score from individual viewpoints in addition to the overall score. Our dataset has five perspectives for a movie: story, cast, direction, visuals, and music. Reviewers are giving a rating on a five-point scale for each perspective. These overall and point-of-view scores were treated as features with one dimension each.

In addition, we also consider how popular the review is and how much it has been viewed. Readers can also vote for a review by pressing the "helpful" button when they find the review useful. In many studies, this vote has been treated as an essential indicator. Review sites have a view count that indicates how difficult the review was to read. View count indicates the amount of interest from non-explicit readers. These values are also included as features since they are essential in determining whether readers believe the review.

3.2 Features of Item

The information about the item reviewed, *i.e.*, the movie, is also expected to affect readers' decision-making. How seriously people read a review may depend on the movie's popularity. The type of review preferred may also differ depending on the genre and the nature of the movie. For example, a concise review will make the reader want to see an action movie. We used the movie's metadata, content, popularity, and reputation as features to represent a movie.

We used running time, release year, and length of title as features to represent general information about a movie. Running time is the movie's length; most movies are about two hours long. The release year was normalized from 0 to 1, since it is a value with a range of about 100 years from 1900 onwards. The length of the movie title was also used as a feature. We suspect that a short and concise title may increase the need for a review because the content cannot be predicted from the title.

Information about the content and quality of the movie is also important. The category tag assigned to a movie is the most straightforward information about the movie's content. In our dataset, there are 15 movie categories: Sci-Fi, Fantasy, Action, Adventure, Animation, and so on. Some of these tags represent a story, and others

represent a technique. Therefore, a single movie often has multiple tags. The category information was expressed as binary features in 15 dimensions.

We additionally used information about the director of the movie. For a movie taken by a veteran filmmaker, the quality and content may be predicted without reading a review. Therefore, we included the total number of movies that the director took in the features. The movie rating is another feature we used to represent the content of a movie. In general, there is a recommended age range for each movie. In Japan, films are classified into four categories: G (for a general audience), PG-12 (parental guidance recommended for children under 12), R15+ (Restricted for under 15), and R18+ (Restricted for under 18). We used it as a 1-dimensional feature taking values one to four.

The popularity of the movie is also an important factor. We used the total number of reviews posted for the movie as a simple popularity measure. The review site we used for analysis allows users to list the movies that they saw, and movies that they want to see, without writing a review. We used these numbers as another popularity measure.

The overall reputation is also considered to contribute to watching the movie. Therefore, we averaged the review ratings of all reviewers for the movie as a 1-dimensional feature to represent the reputation.

3.3 Features of Reviewer

The reviewer who posted the review is also an important piece of information in determining whether the review is useful for decision-making [7]. The reviewer's profile and past posting tendencies are related to the quality of the review. In addition, the reviewer's information may influence whether readers believe the review or not. For example, some reviewers may write incendiary reviews that spur people to action, even though they use the same vocabulary. There may be reviewers with a large fan base (*e.g.*, professional critics who blend into amateurs on the review site). Therefore, we characterize the reviewers into 10-dimensional features by their enthusiasm, the informativeness of their posts, and their authority.

Note that, in this analysis, we did not use reviewers' personal information such as age and gender, but only characteristics related to the review behavior itself. This is because supplying personal information is optional on many review sites, and the reliability of this information cannot be verified. In addition, by using only public review information but not personal information, our model can be used with a wider range of review data.

As the first aspect of a reviewer's characteristics, we focused on their enthusiasm; how much the reviewer has actively contributed to the review site. It has been pointed out that the amount of review experience of a reviewer is related to his or her review ability [15]. As specific features, we used the total number of reviews and the total number of characters as the amount of information in the submitted reviews. On the review sites we analyzed, apart from the number of reviews written, the number of movies the reviewer has seen can be posted as a movie viewing record. Also, if the reviewer has viewed many reviews, then that reviewer has a high interest in this review site. Therefore, we included the number of times a reviewer clicked the "helpful" button for someone else's review in the features.

As for the other characteristics of the reviewers, we focused on informativeness that represents how much information each review of that reviewer had. Specifically, the ratio of reviews with spoiler tags among all reviews and the average number of words in the reviews were used as features. Reviewers can add spoiler tags to each review they write. Reviewers who put a lot of spoiler tags for his or her reviews are more likely to have actually watched the movie all the way through, and mentioned the movie's scenario.

Finally, we quantify the reviewer's authority, *i.e.*, how much the reviewer knows about the movie. For this purpose, we used the grading variance for each viewpoint, the number of impression tags, and the average of the usefulness they received. The variance of rating by viewpoint was used to indicate the reviewer's familiarity with the movie [20]. Some reviewers may focus on only one aspect of a movie to evaluate it. For example, some reviewers may be interested only in the music. Reviews by such reviewers will, for better or worse, influence the reader's decision-making. The variance of the average ratings for each viewpoint was used as an indicator of the reviewer's concern with the viewpoint.

4 ANALYSIS USING REAL DATA

In this study, we used the review data of Yahoo! Movies and the access log of Yahoo! Movies to identify the reviews that contribute to decision-making. This chapter describes the details of the dataset, the actual analysis procedure, and the results.

4.1 Log Data Collection and Cleansing

The dataset used consists of review data from Yahoo! Movies, access logs to each review in Yahoo! Movies, and access logs to the streaming page in GYAO!. The logs were collected over a period of 61 days, from November 1, 2019, onwards and before December 31, 2019.

The access log of the review site consisted of user, movie, review, and visit date and time. This is described for all visitors who accessed the movie during the period. The movie viewing log consists of the user, the movie, and the date and time of the visit. Users visit Gyao! not only from Yahoo! Movies but also from various other sites such as search engines. Therefore, we extracted only the logs where the user with the same user ID actually accessed both sites in a short period of time.

For data cleansing on a per-user basis, we excluded the behavior of users who had never watched a movie on Gyao!. This excludes users who watch movies on other movie streaming sites and users who do not watch movies on the Internet. By limiting the number of users, we can analyze the difference between the actions that encouraged Gyao! users to watch and those that did not.

Finally, we randomly selected 70,000 samples (*i.e.*, a suitable amount for handling and cross-validation). Note that the number of users who moved from Yahoo! Movies to Gyao! is very low compared to the number of users and accesses to these services. The use of review sites is diverse. Some users want to know the reputation of a movie they have seen, and some of the others read reviews to purchase a DVD. Some users watch movies on streaming sites other than Gyao!, even if the logs of users who have never used

Table 2: Performance of the classifier and variants that use limited features

Features	F_1 score
Review itself + reviewer + item	0.771
Review itself + item	**0.786**
Review itself only	0.568

Gyao! were excluded. In this experiment, for data collection reasons, we had to consider that these behaviors were not encouraged.

4.2 Labeling Encouragingness

We defined the encouragingness score $enc(r_k)$ as the ratio of readers who moved to the streaming site after reading a certain review r_k to all readers who read review r_k. The k-th review is expressed as r_k in R, where R is the set of all reviews in the dataset. The encouragingness score $enc(r_k)$ is defined as

$$enc(r_k) = \frac{|U_{watch}(r_k)|}{\sum_{r_p \in R} |U_{watch}(r_p)|} - \frac{|U_{read}(r_k)|}{\sum_{r_p \in R} |U_{read}(r_p)|} \quad (1)$$

where $|U_{read}(r_k)|$ is the number of users who read the review r_k, and $|U_{watch}(r_i)|$ is the number of users who access the streaming site after they read the review r_k. Since $U_{watch}(r_k)$ is a subset of $U_{read}(r_k)$, $encr_k$ falls between 0 and 1.

The value of $enc(r_k)$ is continuous. There are two ways to use such a value as a correct answer label. One way is to set a threshold and divide it to 0 or 1. The other way is to split using a fixed percentile. In this analysis, we consider the top ten percent as correct answers and the other 90 percent as negative examples. For the random forest training, the same number of samples were used from correct and negative examples.

4.3 Feature Analysis by Classification and Cross-Validation

We classified these reviews into encouraging reviews and other reviews using random forest. As the implementation of random forest classifier, we used the existing machine learning library called scikit-learn[4]. First, we made two variant methods; one uses only features of the review itself, while the other uses review and item features. The performance of each method was calculated by five-fold cross-validation. Classification and cross-validation were done using the initial parameters of scikit-learn. The number of trees is 300, and the maximum depth of a tree is 10.

The classification result is shown in table 2. The best performing classifier was the one that used the features of the review itself and the item, with an F_1 score of 0.786. On the contrary, the model using all the features showed decreased accuracy. This indicates that the features associated with the movie have a significant impact on the classification performance of the review's encouragingness.

Next, the importance of each feature of the classifier using all the features was tabulated for each element of interest. In scikit-learn, which was used in this experiment, the importance of a feature in a random forest is the average of the importance of the nodes of each decision tree that composes the forest for each feature. In a random forest, multiple short decision trees are created, and not all features are necessarily used in all trees. Let f_i be the feature value of the i-th dimension of the feature vector. For a node n_i to be classified by feature f_i in a tree t_j, the probability of reaching it is expressed as $p(n_i)$ (this probability will be used as a weight). Since the decision tree is a binary tree, the child nodes connected to a node n_i are divided into right and left, and denoted as (n_{ir}) and n_{il}. The importance $imp_{\text{tree}}(n_i, t_j)$ of feature n_i in tree t_j can be represented as

$$imp_{\text{tree}}(n_i, t_j) = p(n_i)g(n_j) - p(n_{ir})g(n_{ir}) - p(n_{ir})g(n_{ir}) \quad (2)$$

where impurity is represented by $g(n_i)$. For all the trees T used in that random forest classifier, the importance of a feature f_i is

$$imp_{\text{forest}}(f_i) = \frac{1}{|T|} \sum_{t_j \in T} imp_{\text{tree}}(n_i, t_j), \quad (3)$$

i.e., the average of the importance of the feature in all trees.

Table 3: Importance for each set of features.

Type	Hypothesis	Importance
Review itself	Amount of information	0.056
(Total: 0.305)	Readability	0.048
	Review content	0.137
	Popurality	0.064
Item	Metadata	0.039
(Total: 0.425)	Item content	0.117
	Popurality	0.210
	Reputation	0.059
Reviewer	Enthusiasm	0.138
(Total: 0.273)	Informativeness	0.052
	Authority	0.083
Total		1.000

The importance of each set of features is summarized in Table 3. In terms of the contribution of each set of features, features associated with the movie have the highest importance. The total importance of the item features was 0.42. The importance of the review content was the next highest at 0.30, and the reviewer-related features were the lowest at 0.27. From the most important items, we were able to deduce that the movie's popularity is especially encouraged by the readers' movie-watching behavior. The other important features were the reviewer's enthusiasm and the content of the review.

5 SUBJECT EXPERIMENT

In order to confirm whether the findings obtained by the feature analysis are useful in the real world, we conducted a subject experiment. In this experiment, we check whether the classification results by Random Forest differ from the actual human feeling. For this purpose, we collected reviews that were not used in training (*i.e.*, reviews for movies that were not distributed on the streaming site), and classified them by whether they encourage readers to watch the reviewed movie. Experiment participants were asked in a

[4]scikit-learn:https://scikit-learn.org/

questionnaire whether reading the review had actually made them want to see the movie.

5.1 Experimental Setting

We conducted a questionnaire-based study to analyze whether the reviews that were determined by the classifier to be encouraging of movie-watching behavior are actually considered encouraging to participants. For the experiment, we selected the ten movies with the highest box-office revenue in Japan in 2019. We used the trained classifier to estimate the encouragingness of all the reviews posted for each of these movies. We sampled the reviews by the confidence score of the random forest classification: top five and bottom five.

The participants in the experiment were six university students. The reviews extracted by the classifier were presented for participants in random order. Participants were asked to indicate for each review whether they felt encouraged to watch that movie when reading the review. They were asked to rate each review on a four-point scale, from 1 (not encouraged) to 4 (well encouraged). The review was presented in a stand-alone text format. In other words, no metadata about the film and no information about the review's author was provided.

We compared the classification results of four methods: the proposed method that includes all features, the variant method that uses item information and reviews, the method that uses only reviews itself, and a baseline method. The baseline method randomly selects and ranks ten reviews from the entire set of reviews of a movie without using any classifier.

5.2 Result

The score assigned by humans for each method is shown in Table 4. Reviews estimated to be encouraging by the classifiers were rated as more facilitative for humans than other reviews. These differences were significant at $p < 0.01$ as a result of validation by Student's t test. For the randomly selected reviews, there was no difference between the top and bottom, as a matter of course.

Throughout, the average of reviews' encouragingness score was 2.45 on our four-point scale. It suggests that most reviews on the review site resulted in an evaluation of "slightly encouraged to see the movie". Focusing on the values, the method that best estimated the encouraging reviews was the classifier that only used features of the item and the review itself. The difference between the top and bottom is also the highest at 0.88. This indicates that the classification results by this method and the degree of encouredingness felt by the human are close. In contrast, the classifier using the author's characteristics disagreed with participants.

6 DISCUSSION

Overall, we were able to estimate whether the reader would actually go to the streaming site after reading the review by using the features surrounding the review. Through the experiment where we show the reviews estimated encouraging to the participants, the classification results and the participants' opinions generally agreed. These results suggest that our method is able to correctly extract reviews that encourage people's viewing behavior.

Let us discuss the features of the movies that were judged to be the most effective. From the comparison of classifiers (see table 2) and the importance of the features (see table 3), it was found that the features associated with the movie contributed greatly to the estimation of whether or not the movie encouraged movie-watching behavior. This indicates that whether a person wants to watch a movie is more influenced by the reviewed movie itself than by the text of the review.

Among the features of the movie, the popularity of the movie had the greatest effect. In other words, if a movie is popular, users will feel compelled to watch that movie regardless of the content of the reviews. The popularity of a movie has a larger impact than other movie features (*e.g.*, content, metadata). It is possible that people judge whether to see a movie by its popularity, not by its content. Among the features associated with a movie, the reputation information of a movie has a relatively small impact. Even if a movie is heavily criticized in reviews, people may still watch it if it is a famous movie.

Here we need to consider the way movie review sites are used. People may first decide which movie to watch and then go to the review site to check the reviews. Or, they may somehow find the movie they want to see by looking at reviews for various movies and then go to the streaming site if there is a link. Since the usage of review sites is different for each person, additional analysis which focuses on individual users is needed in the future.

Next, we discuss the features of the review contents. Even with a classifier that uses only the features inherent in the review for classification, we were able to extract reviews that entailed viewing with an accuracy of $F = 0.57$. From the importance of the features shown in table 1, the features related to review content had the greatest effect for encouraging movie-watching. The features with the highest importance were the grades of each viewpoint, followed by the topic of the review text. This indicates that users may pay attention to the viewpoint when reading reviews. For example, a user who is interested in music might look at the grades related to music, and then read the sentences related to music in the review text.

In addition, we found that the amount of information in the review did not have much effect on movie viewing behavior. From this result, we can deduce some tips for writing reviews that encourage people. That is, it is futile to spend the effort to write lengthy, informative reviews. Reviewers should pay attention to the viewpoints, and review the film from each viewpoint.

Finally, we also found that there is still room for improvement in the method. The label of correct answers seems unnatural. The negative examples in the behavioral logs may include people who have already seen the movie or users of another streaming site. In addition, the dataset contained movies that were seen by many people and movies that only a few people saw. These factors may give some biases to our dataset. The dataset itself needs to be analyzed and cleansed. We analyzed only two months' worth of logs (in consideration of privacy protection). Therefore, there is a possibility that major movies that are released during that period may have affected the results. Long-term analysis is desired in the future.

7 CONCLUSION

Using a large-scale access log of an actual online movie review site, we analyzed the characteristics of the reviews that encouraged

Table 4: Result of the subject experiment if the review encourages the participant to watch the reviewed movie released in 2019 (* means the statistically significant difference of t test; * $p < 0.05$, ** $p < 0.01$)

title	All features		Review itself + item		Review itself only		Random	
	Top 5	Bottom 5	Top 5	Bottom 5	Top 5	bottom 5	Top 5	Bottom 5
Kingdom	2.90*	1.70	3.50**	1.20	3.40*	2.50	2.90	2.70
Toy Story 4	2.90	2.70	1.40	2.40	1.70	1.60	1.40	1.60
Detective Conan	1.90	2.10	2.60*	1.50	2.60*	1.40	2.20	1.90
ONE PIECE: STAMPEDE	2.70**	1.90	3.20**	1.70	3.10*	2.20	3.10	2.40
Weathering with You	2.50**	1.80	2.80	2.40	3.60**	2.10	3.50	3.60
Avengers: Endgame	2.40	2.00	2.50*	1.50	2.40	2.20	2.20	2.90
Aladdin (2019)	3.10**	1.40	3.20*	2.20	2.60	3.00	2.50	2.60
Star Wars Ep. 9	2.10	1.70	3.20**	1.70	2.60	2.50	2.10	2.00
Frozen II	2.70*	1.60	3.00	2.60	3.30	3.00	2.90*	2.30
The Lion King (2019)	2.20	2.60	3.00	2.40	3.00*	2.00	3.20*	2.60
average	2.54**	1.95	2.84**	1.96	2.83**	2.25	2.60	2.46

readers to watch a movie after reading a review of the movie. We focused on the users' page transitions from the review site to the movie streaming site after they read the review. For the analysis, we used a random forest classifier trained to determine whether a review caused a movie-watching behavior. We conducted feature importance-based analysis using three types of features; features of the review itself, features about the item, and features of the reviewer. We used the actual access log of Yahoo! Movies Japan (one of the biggest movie review sites in Japan) and Gyao! (a movie streaming site operated by Yahoo! Japan). Through the cross-validation experiment, the classifier was able to classify encouraging reviews. In particular, features related to the movie's popularity were shown to affect the encouragement of people's movie viewing. A subject experiment confirmed that these features contribute to the review's usefulness.

The contribution of this study is that we made it possible to discover encouraging reviews from actual access log, and that the popularity of the movie and the author's review experience affected the readers' decision-making process. As future work, we need to evaluate whether classified reviews actually encouraged users. We will be able to conduct A/B testing-based user experiments using our results. Application is also an important technical issue. For example, when a reviewer posts a new review, the system could display advice that says, "Your review will be better if you change here ." These would be of great help in improving the quality of review sites.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grants Number 18K18161, 21H03775, and 18H03243.

REFERENCES

[1] Einar Bjering, Lars Jaakko Havro, and Øystein Moen. 2015. An empirical investigation of self-selection bias and factors influencing review helpfulness. *International Journal of Business and Management* 10, 7 (2015), 16.

[2] Don Charlett, Ron Garland, Norman Marr, et al. 1995. How damaging is negative word of mouth. *Marketing Bulletin* 6, 1 (1995), 42–50.

[3] Pei-Yu Chen, Samita Dhanasobhon, and Michael Smith. 2008. All Reviews are not created equal: The disaggregate impact of reviews on sales on Amazon. com. In *Carnegie Mellon University Working Paper*.

[4] Christy MK Cheung, Matthew KO Lee, and Neil Rabjohn. 2008. The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities. *Internet research* 18, 3 (2008), 229–247.

[5] Judith A. Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43, 3 (2006), 345–354.

[6] Alton Chua and Snehasish Banerjee. 2014. Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth. *Journal of the Association for Information Science and Technology* 66 (06 2014). https://doi.org/10.1002/asi.23180

[7] L. Connors, S. M. Mudambi, and D. Schuff. 2011. Is It the Review or the Reviewer? a Multi-Method Approach to Determine the Antecedents of Online Review Helpfulness. In *2011 44th Hawaii International Conference on System Sciences*. 1–10. https://doi.org/10.1109/HICSS.2011.260

[8] Chris Forman, Anindya Ghose, and Batia Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information systems research* 19, 3 (2008), 291–313.

[9] Yuanyuan Hao, Qiang Ye, Yijun Li, and Zhuo Cheng. 2010. How Does the Valence of Online Consumer Reviews Matter in Consumer Decision Making? Differences between Search Goods and Experience Goods. 1–10. https://doi.org/10.1109/HICSS.2010.455

[10] Hong Hong, Di Xu, G Alan Wang, and Weiguo Fan. 2017. Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems* 102 (2017), 1–11.

[11] Liqiang Huang, Chuan-Hoo Tan, Weiling Ke, and Kwok-Kee Wei. 2013. Comprehension and assessment of product reviews: A review-product congruity proposition. *Journal of Management Information Systems* 30, 3 (2013), 311–343.

[12] Zhiwei Liu and Sangwon Park. 2015. What makes a useful online review? Implication for travel product websites. *Tourism Management* 47 (2015), 140–151.

[13] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th International Conference on World Wide Web*. 691–700.

[14] Jing Luan, Zhong Yao, FuTao Zhao, and Hao Liu. 2016. Search product and experience product online reviews: An eye-tracking study on consumers' review search behavior. *Computers in Human Behavior* 65 (2016), 420 – 430. https://doi.org/10.1016/j.chb.2016.08.037

[15] Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*. 897–908.

[16] Susan M Mudambi and David Schuff. 2010. What makes a helpful review? A study of customer reviews on Amazon. com. *MIS quarterly* 34, 1 (2010), 185–200.

[17] Phillip Nelson. 1974. Advertising as information. *Journal of political economy* 82, 4 (1974), 729–754.

[18] Yue Pan and Jason Q Zhang. 2011. Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of Retailing* 87, 4 (2011), 598–612.

[19] Pradeep Racherla and Wesley Friske. 2012. Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications* 11, 6 (2012), 548–559.

[20] Yoshiyuki Shoji, Makoto P. Kato, and Katsumi Tanaka. 2014. *Can Diversity Improve Credibility of User Review Data?* Springer International Publishing,

Cham, 244–258. https://doi.org/10.1007/978-3-319-13734-6_17

[21] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems*. 1–8.

[22] Y. Zhou and S. Yang. 2019. Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews. *IEEE Access* 7 (2019), 27769–27780. https://doi.org/10.1109/ACCESS.2019.2901472