Identifying and Analyzing Obscure Venues Using Obscure Words in User-provided Reviews

Masaharu Hirota
Faculty of Informatics
Okayama University of Science
Okayama-shi, Okayama
Email: hirota@mis.ous.ac.jp

Jhih-Yu Lin
Graduate school of System Design
Tokyo Metropolitan University
Hino-shi, Tokyo
Email: lin-jhihyu@ed.tmu.ac.jp

Masaki Endo
Division of Core Manufacturing
Polytechnic University
Kodaira-shi, Tokyo
Email: endou@uitec.ac.jp

Hiroshi Ishikawa
Graduate school of System Design
Tokyo Metropolitan University
Hino-shi, Tokyo
Email: ishikawa-hiroshi@tmu.ac.jp

***Abstract*—When sightseeing, many people visit different places such as restaurants, hotels, and tourist spots. Some of these venues, while worthwhile, are considered obscure, secret, not well-known, or having little popularity. Their extraction and recommendation are vital to improving the satisfaction of tourists. This research proposes a method for discovering obscure venues using classifiers for identifying reviews, including obscure impressions. To achieve this goal, in this research, a model was developed to classify venues as obscure or not obscure using reviews with language indicating their obscurity. In addition, we compare various methods for generating feature vectors and the models for classification. This research also analyzes the differences among venues perceived by reviewers as being obscure. We demonstrate the performance of the proposed approach by indicating that the posting destination of obscure reviews differs for each user.**

Keywords–Tourism information; Text classification; Support Vector Machine; Review Analysis.

I. INTRODUCTION

A considerably shorter pre-version of this paper has already been published in [1].

In recent years, it has become commonplace for many people to give their opinions and impressions regarding several spots as tourist spots, hotels, and restaurants, on review websites such as Yelp [2], Expedia [3], and TripAdvisor [4]. In this paper, we call such spots venues. Reviews written about venues describe information regarding the venues themselves and the impressions to them and behaviors of the users. Such reviews are useful for travel planning, obtaining information on travel destinations, tourist behavior, and visitor impressions of popular tourist spots. Therefore, many studies have extracted tourism information from user-provided reviews [5][6].

Some venues are obscure, secret, not well-known, or having little popularity. Despite not being popular, such venues may be well-regarded by visitors. In this paper, these are collectively called "obscure". Because some obscure venues can lead to improved tourist satisfaction and the acquisition of repeat visitors, some methods for describing obscure venues and recommending them to tourists have been proposed [7] [8]. Definitions regarding obscure venues have been proposed in such studies. Studies on this subject commonly define an obscure venue as one in which the visibility for tourists is low, but the value is high. For example, the authors in [7] defined obscure spots as less known, but still worth visiting, and extracted such spots. Also, [9] extracted hidden tourist spots with low popularity but a high level of satisfaction. However, precisely identifying obscure venues is difficult because the places that people feel are obscure depends on their own personality.

In this research, we identify obscure venues from review sites, and the proposed approach focuses on words in the text of the venue reviews. This research then extracts obscure reviews without directly defining obscure to accommodate the fact that the impression of a venue differs among different people. For this research, we regard a venue with many reviews written about the impression of its obscurity as an obscure venue (hereinafter referred to as "obscure review". Also, we call other reviews "non-obscure review").

This research extracted such reviews from all reviews on a particular venue. In this paper, a review is defined as an obscure review if its text contains terms related to "obscure" (hereinafter referred to as "obscure words"). If the ratio of reviews of a venue that includes obscure words accounts for the majority, the venue is defined as obscure.

The aim of this research is the identification of obscure venues using user-provided reviews that include obscure words. However, in most cases, the number of reviews on a venue is small. Because an obscure venue might be less well-known by people even if worthwhile, there will be few reviews for such venues. Also, few reviews obtain obscure words. As a result, the number of reviews to be classified as obscure is insufficient for identification of obscure venues. Moreover, it is unrealistic to define all expressions related to the word obscure. Therefore, to extract obscure reviews that do not include obscure words but rather the description of an obscure venue, this research applies the classification model of the representation of contents of a review as obscure or not, regardless of whether a review contains an obscure word. Reviews that do not contain obscure words were classified using the model, and the classifier was evaluated using a dataset of reviews submitted by users.

Moreover, different reviewers have posted various reviews

on different venues, and the criteria by which a venue is considered obscure differs according to the reviewer. Therefore, this research revealed that the reviewer who posts an obscure review for each venue is different. As a result, this research examined the efficiency of the proposed approach in identifying obscure venues using the obscure-word based classifier without a direct definition of the term obscure.

A summary of contributions from this research is as follows.

- We design a new approach for identifying obscure venues using user-provided reviews.
- We propose a classifier for identifying obscure reviews without the obscure words.
- We analyze the posting destination of obscure reviews differently for each user.

The remainder of this paper is organized as follows. Section II presents previous studies related to this topic. Section III describes our proposed method for the development of a classifier for discovering obscure reviews by using obscure words and the identification of obscure venues. Section IV describes the experiments evaluating our proposed method using the Yelp dataset. Section V describes an analysis of the hypothesis that an obscure venue is perceived differently for each user and discuss the extracted obscure reviews and venues. Section VI provides some concluding remarks along with a discussion of results and areas of future work.

II. Related Works

The main aim of our research was to find obscure venues for tourism analysis using user-provided reviews posted to social media sites. This section introduces the related studies published in the area of analysis of tourism information using reviews and extracting obscure venues. Also, vectorizing documents is an essential procedure for review analysis, because the performance of vectorization has massive effects on the classification of them. Therefore, we describe the related studies of document vectorization.

A. Analysis for tourism using reviews

Research has been conducted on the extraction of tourism information through user-generated content on social media sites. Also, extracting helpful or useful information from text data like reviews and blogs is one of the research tools used to analyze reviews. Our proposed research on extracting obscure venues from reviews is related to the analysis of reviews for recommendations and the analysis of tourism information.

[10] analyzed factors affecting the perceived usefulness of reviews to findings contributing to tourism marketers. [11] predicted where memorable is the travel destination using the user-generated photographs in blogs. [12] proposed a method for identifying dimensions of satisfaction using an unsupervised learning algorithm with numerical and textual information from user-generated online reviews, and analyzed the multiple factors contributing to consumer satisfaction. [13] predicted how helpful a review is and presented a list of ranked reviews based on an evaluation. [14] proposed a method for detecting reviews that reliably predict foodborne illnesses using review classification. [15] analyzed online review to identify insights through a case study, and found them. For example, overall review star rating correlates well with the sentiment scores for both the title and the full content of the online reviews. [16] proposed a method for detecting the topic of phrases in helpful recommending reviews. [17] proposed a method for aspect-based opinion mining of tourism reviews to classify them into negative or positive aspects. [18] proposed an approach for sentiment classification of online hotel booking opinions using a dependency tree structure. [19] investigated the valence of online reviews and modeled them with hotel attributes and performance. [20] analyzed the online reviewer profile, and exposed its image can significantly enhance consumers evaluation of review helpfulness. [21] concluded that sentiment analysis plays an important role in the analysis of tourism reviews and summarize their studies.

These studies analyzed user-provided reviews on social media sites for improving sightseeing satisfaction. This paper tackles the analysis of user perception of obscure venues based on reviews.

B. Extracting obscure venues from social media sites

Studies have been conducted on extracting obscure venues and tourist spots from social media sites. Because obscure spots are expected to spread tourists to other tourist spots and improve the satisfaction of the tourism experience, some studies extracting posts on such spots have been conducted.

[7] proposed a method for evaluating sightseeing spots that are less well-known but are worth visiting. [8] defined the term obscure to indicate spots that are not famous but have high evaluations, and extracted such spots based on name recognition and user evaluations. [9] proposed a method for providing tourism information on hidden spots for increasing tourism satisfaction. [22] extracted hot and cold spots based on a spatial analysis of user-generated content to extract knowledge of tourist behaviors. [23] proposed a method for less-known tourist attractions by using a clustering algorithm from geo-tagged photographs on Flickr.

This study used a classifier to extract obscure venues using reviews that include the word obscure to comprehensively deal with familiarity, popularity, and attractiveness. The main characteristic of this research is the extraction of sightseeing spots recognized by reviewers as obscure venues by using the classifier.

C. Document vectorization

Various methods have been proposed to vectorize documents. The traditional approach for vectorizing documents is some hand-craft features such as Term Frequency-Inverse Document Frequency (TFIDF), bag-of-words, and n-grams. Also, unsupervised representation learning have been used [24], [25], [26]. However, in recent years, pre-trained deep language representation model has been highly successful in the domain of Natural Language Processing (NLP) [27], [28], [29], [30]. Especially, Bidirectional Encoder Representations from Transformers (BERT) [27] achieved state-of-the-art performance on various NLP tasks.

To generate a document vector using BERT, the most commonly used approach is to average the BERT output layer or by using the output of the first token (the [CLS] token). However, this approach has been pointed out as unsuitable by [31]. Sentence-Bert [31] (SBERT) is a model that BERT [27] to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. Therefore, in this paper, we use SBERT for generating a document vector.

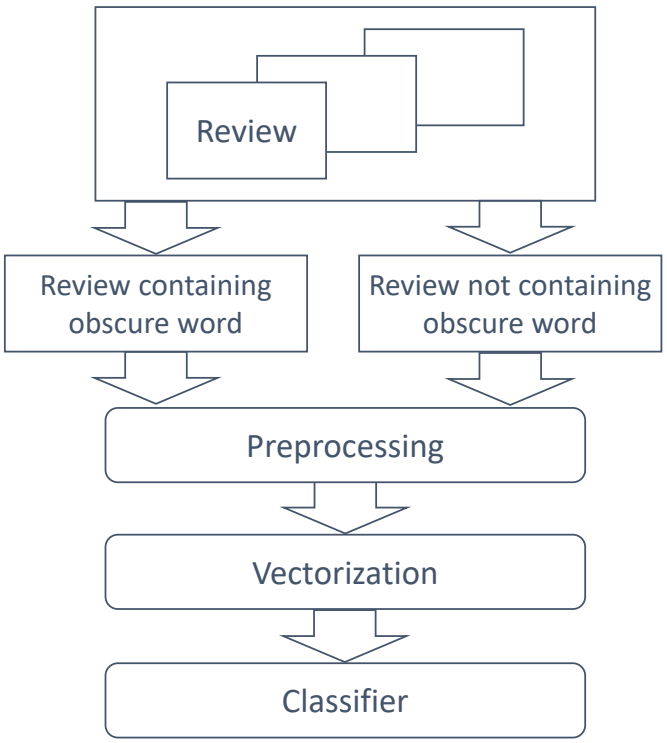

Figure 1. Overview of classifier for extracting obscure reviews using obscure words.

TABLE I. OBSCURE WORDS.

secret spot	secret place
best kept secret	best-kept secret
well-kept secret	well kept secret
local secret	obscure spot
hidden spot	hidden place
little known	little-known
good out of the way	

III. Proposed Method

In this section, we describe our proposed method for discovering obscure venues using obscure words and classification algorithms from user-provided reviews.

This research extracted reviews including obscure words and generated a classifier for both obscure and non-obscure reviews. We indicate an overview of our proposed classifier in Figure 1. First, we extract obscure and non-obscure reviews from reviews. Next, we apply the preprocessing method for the reviews. Next, we apply a vectorization method to generate a document vector. Finally, we create a model of the classifier using a vector to classify a review as obscure or not.

After this process, the classifier is applied to all reviews on a venue, and the venue is classified as obscure or non-obscure based on the reviews classified as obscure.

A. Obscure words

This section explains obscure words that we used for extracting obscure reviews.

In this research, obscure words are used to identify obscure venues from all reviews in a venue. This research defined 13 obscure words, as shown in Table I. The criterion for selecting obscure words is to select an English phrase manually that seems to represent a word indicating obscurity, and an expression that has no meaning other than obscurity.

However, these words do not cover all words expressing user perceptions of obscurity. Also, for example, phrases such as "little well known" can assume word choices and various spelling variations. Preparing all those phrases or words included in reviews is not realistic. However, it is desirable to extract obscure reviews from all of them that contain unknown obscure phrases or do not include those phrases. Therefore, we conduct supervised learning using obscure reviews including these words to discover obscure reviews not including them.

B. Preprocessing

This section describes the preprocessing applied to vectorize the reviews for machine learning.

First, reviews written in English were extracted from all reviews. In this paper, to detect the language of the texts we applied langdetect [32] to them.

Also, we extract reviews where the text has more than 30 words. This reason is because the classification is difficult when the number of words is small.

The texts from the extracted reviews were converted into lower-case texts. Next, we apply stop-word elimination and stemming to each word. This research defined 319 stop words, such as "the" and "and," which are commonly used in sentences.

C. Vectorization

Next, the preprocessed reviews were vectorized for determining what words in reviews might be more efficient for extracting obscure reviews. In this paper, we tried two vectorization methods. First, is TFIDF, which is one of the major hand-craft features. The other is SBERT, which is a pre-trained deep language representation model.

1) TFIDF: First, TFIDF were applied to the texts.

In this paper, we calculated the TFIDF of each word t in review r. The term frequency $tf(t,d)$ and inverse document frequency $idf(t,D)$ are calculated using the follow equations:

$$tf(t,r) = \frac{f_{t,r}}{\sum_{t \in r} f_{t,r}} \tag{1}$$

$$idf(t,R) = \log \frac{|R|}{|\{r \in R : t \in r\}|} \tag{2}$$

where the number of reviews is $|R|$, and $f_{t,r}$ is the number of occurrences of word t in review r.

Then, the TFIDF of each word t in review r in reviews R is calculated through the following equation:

$$tfidf(t,r,R) = tf(t,r) \times idf(t,r) \tag{3}$$

2) SBERT: SBERT is a modification of the pre-trained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using a similarity function.

SBERT uses the output of CLS-token or all output vectors from BERT. First, this model applies the pooling operation to the vector. In this paper, we adopt all output vectors and mean pooling. Also, to fine-tune BERT, SBERT used siamese networks to update the weights and the objective function is the following equation:

$$\omega = softmax(W_t(u, v, \|u - v\|)) \tag{4}$$

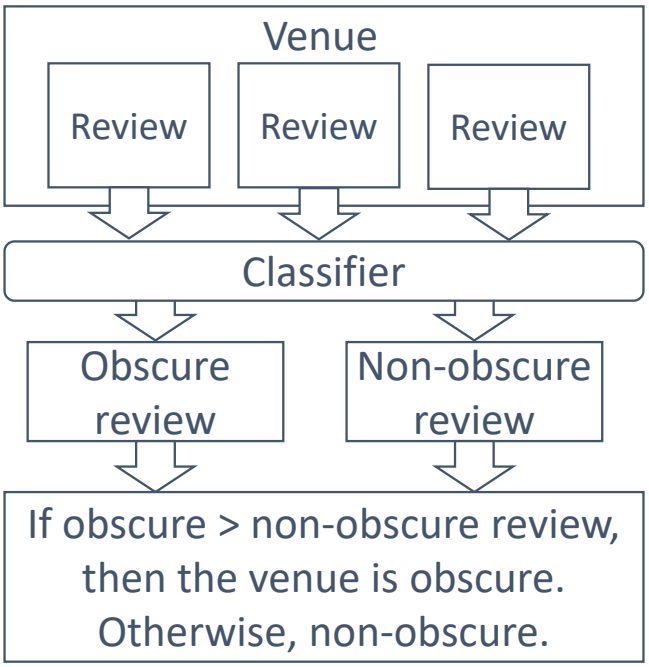

Figure 2. Overview of procedure for identification of obscure venues using obscure and non-obscure reviews.

Here, n is the dimension of the sentence embeddings and k is the number of labels, W is the weights of siamese network.

In this paper, we used the pre-trained model of SBERT using NLI models [31], in which this model was generated using the combination of two datasets [33] [34]. Next, the preprocessed reviews were vectorized for determining what words in reviews might be more efficient for extracting obscure reviews.

D. Classification of obscure reviews

In this section, we describe the procedure for generating a classification model of reviews regardless of whether they are obscure reviews. Our method proposed in this research identifies obscure venues using obscure reviews even if the review does not include obscure words. Therefore, our proposed method creates a classifier for identifying such reviews that do not include obscure words but when their content represents an obscure venue.

A method is proposed to classify the reviews into obscure or non-obscure reviews. In this research, we apply a binary classification method using vectors generated as described in Section III-C. The first class is thus obscure reviews, which consists of reviews that contain an obscure word. The other class is non-obscure reviews, which consists of reviews that do not contain an obscure word.

This research used three binary classification methods to classify reviews as obscure or not obscure: Support Vector Machine (SVM) [35], Random Forests (RF) [36] and LightGBM [37].

E. Identification of obscure venue

Herein, we describe how to find obscure venues using a classifier. Figure 2 shows an overview of the procedure for the identification of an obscure venue. We collect all review texts of a venue and apply the classifier described in Section III-D to the reviews. Finally, we count the reviews classified as obscure or non-obscure reviews of a venue. As a result, this research regards an obscure venue as one in which the percentage of obscure venues is greater than the threshold. In this paper, when the ratio of reviews classified as obscure among all reviews on a venue is larger than half, the venue is considered obscure, otherwise, it is non-obscure.

IV. Experiments of classification performance

In this paper, we evaluate the performance of our proposed method through an evaluation experiment based on classification. We describe the experimental conditions of the dataset and the evaluation criteria. Also, we describe our experiments conducted for the evaluation of obscure review discovery.

A. Dataset

Herein, we describe the dataset used for this experiment. We used the Yelp Dataset Challenge (round 13) [38], which includes 192,609 venues and 6,685,900 reviews which were written by 1,637,138 users. After we applied the preprocessing procedure as described in III-B, the number of reviews, venue, and users is 518,8614, 165,060 and 602,988, respectively.

This research comprises 1,780 reviews that mention an obscure word at least once. Table II shows the number reviews containing each obscure words. Here, we replaced the name of the venue into "@" to anonymize it. About 45% of reviews in the table contain the word "best kept secret". Therefore, this word is a general phrase for representing obscure venues. However, this table shows various words representing obscure venues are used. We used these reviews to generate a classifier for identifying a review as obscure or not. Also, we prepared the same number of randomly selected reviews from reviews which do not contain the obscure words.

The reviews with and without obscure words were randomly split into a ratio of 4:1 for training and testing data. As a result, the number of training data is 2,848 and testing data is 712.

B. Evaluation criteria

We used the following widely used performance measures for classification: Accuracy, Recall, Precision, and F-measure. To calculate them, we exploited the concepts of True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN), which shown in Table III. TP is the number of obscure reviews that are predicted as obscure. TN is the number of non-obscure reviews that are predicted as non-obscure. FP is the number of non-obscure reviews that are predicted as obscure. FN is the number of obscure reviews that are predicted as non-obscure. Using them, Accuracy, Recall, Precision, and F-measure are calculated as the following equations.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + TN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

TABLE II. THE NUMBER OF REVIEWS FOR EACH OBSCURE WORD AND EXAMPLES OF A PART OF REVIEWS

Obscure word	The number of reviews	Example
secret spot	106	I only gave 3 stars because I don't want you blowing up my secret spot!
secret place	112	It is a secret place that not even all locals know about, and the pizza is great.
best kept secret	789	As others have said, this place is one of the best kept secrets.
best-kept secret	71	@ it's a best-kept secret you only share with your close friends.
well kept secret	42	Overall, @ has been a well kept secret amongst those in the know for a long time.
well-kept secret	96	We enjoy coming here nonetheless. Maybe it's just a well-kept secret.
local secret	39	Even though many people knew about it, it still seemed like a local secret.
obscure spot	4	One of our favorite relaxed but obscure spots with the decor of an opium den slash western saloon.
hidden spot	155	It's in such a hidden spot you wouldn't know it was there unless you looked it up or saw people walking out of a hallway with a pizza box!
hidden place	124	I've been wanting to try this little hidden place for over a year now and finally found the time.
little known	188	If it's not on someone's list of high-quality, yet little known local-area sports bar destinations it should be.
little-known	43	I'd say it's the best little-known hard dip ice cream place in town.
good out of the way	11	First, let me get the good out of the way. The kids who got my order were nice, and the restaurant was clean.
sum	1,780	

TABLE III. CONFUSION MATRIX.

		Predict	
		Positive	Negative
Correct	Positive	TN	FN
	Negative	FP	TP

TABLE IV. EVALUATION RESULT: ACCURACY.

	RF	SVM	LightGBM
TFIDF	0.74	0.74	0.74
SBERT	0.75	0.77	0.73
Ave.	0.75	0.76	0.74

TABLE V. EVALUATION RESULT: PRECISION.

		RF	SVM	LightGBM
TFIDF	Obscure	0.76	0.76	0.76
	Non-obscure	0.73	0.74	0.74
	Ave.	0.75	0.75	0.75
SBERT	Obscure	0.78	0.82	0.75
	Non-obscure	0.73	0.74	0.73
	Ave.	0.76	0.78	0.74

TABLE VI. EVALUATION RESULT: RECALL.

		RF	SVM	LightGBM
TFIDF	Obscure	0.72	0.72	0.73
	Non-obscure	0.76	0.77	0.77
	Ave.	0.74	0.75	0.75
SBERT	Obscure	0.71	0.72	0.72
	Non-obscure	0.80	0.84	0.76
	Ave.	0.76	0.78	0.78

C. Experimental conditions

This section describes the procedure used for the creation of classifiers for obscure reviews.

This experiment used a Gaussian kernel for the SVM kernel function and entropy and Gini impurity for a split of nodes in Random Forest. In addition, the hyperparameters of those methods were searched using Optuna [39] with five cross-validations, which is a software for automatically optimizing hyperparameters. We used the parameters with the highest accuracy measured through this experiment. In addition, we used the Python software scikit-learn [40] for the implementation of the SVM, RF, TFIDF, and evaluation criteria in the following experiments. Also, we used [41] for the implementation of the LightGBM.

D. Evaluation results

In this section, we describe and discuss the evaluation results of classifying reviews into obscure or non-obscure reviews.

Table IV shows the evaluation results of the classification of obscure reviews through the procedure described above using accuracy. Also, Tables V, VI, and VII show the evaluation results of the classification of obscure reviews through the procedure described above using f-measure, precision, and recall. In those tables, "Obscure review" shows the reviews that include an obscure word, whereas "Non-obscure review" shows reviews that do not include an obscure word.

Also, in Tables IV, V, VI, and VII, comparing TFIDF and SBERT used for document vectorization, the evaluation scores of the SBERT is better than TFIDF in most cases. This reason is that the procedure for generating TFIDF is a simple way and does not consider the context and meaning of sentences, but SBERT uses a complex model considering them and can generate better feature vector. Also, the evaluation score of the combination of RF and LightGBM with TFIDF has often better performance than SBERT. As described in Section III-C1, a dimension in the vector generated by TFIDF shows the degree of appearance of one word in one document. On the other hand, the vector of SBERT is generated using neural networks and one vector does not have a specific role. Also, RF and LightGBM have functions for feature engineering such as feature bagging and exclusive feature bundling. Therefore, we think that those methods choice better dimensions from document vectors by themselves and showed better performance. However, SBERT generated a better feature vector in the overall evaluation.

Comparing the results shown in Tables VII, V, and VI for obscure and non-obscure reviews, the evaluation scores of the non-obscure reviews are lower than those of the obscure reviews. In particular, there is a vast difference between both scores regarding the recall rate. The evaluation score is achieved because reviews with an obscure word are misclassified as non-obscure in certain cases because the number of reviews in the training dataset is unbalanced. However, the purpose of this research is to identify obscure venues using extracted obscure reviews. As shown in Table V, the precision of the obscure reviews was 0.82 (the combination of SBERT and SVM), which shows that it is rare for a classifier to misclassify the content of reviews unrelated to obscurity.

In Table IV, the best combination of feature and classification methods is SVM and SBERT in almost cases. However,

TABLE VII. EVALUATION RESULT: F-MEASURE.

		RF	SVM	LightGBM
TFIDF	Obscure	0.74	0.75	0.75
	Non-obscure	0.75	0.74	0.75
	Ave.	0.75	0.75	0.75
SBERT	Obscure	0.75	0.76	0.74
	Non-obscure	0.76	0.79	0.74
	Ave.	0.76	0.78	0.74

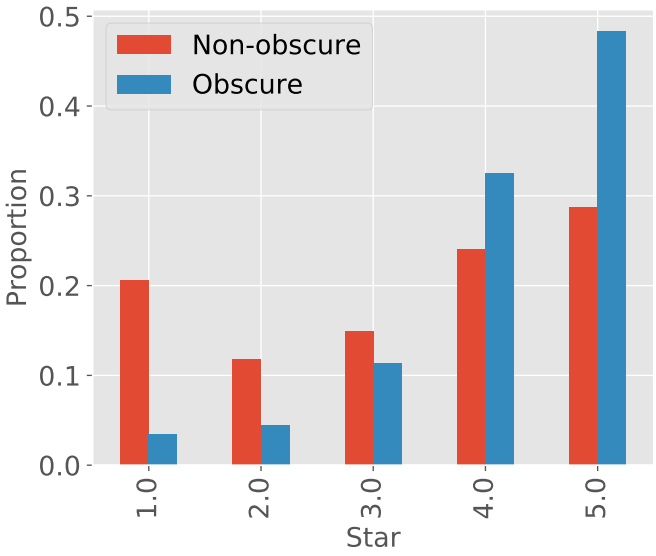

Figure 3. Distributions of proportion of stars on obscure and non-obscure reviews.

the difference in performance with other combinations is small. These results show that various document vectorization and classification methods can classify reviews into obscure and non-obscure. Therefore, our approach which uses the reviews containing obscure words to discover obscure reviews is effective.

E. Analysis of stars in obscure and non-obscure reviews

This section describes and discusses the difference of the stars in Yelp between obscure and non-obscure reviews. Here, the star represents the evaluation score, in which a reviewer evaluates a venue on a scale of 1 to 5 (1 = bad and 5 = good).

In general, we think that reviewers who wrote obscure reviews are considered to have a positive evaluation to the venue. Therefore, we can assume that the stars of obscure reviews are high. On the other hand, in the case of non-obscure, the reviewer wrote not only positive ratings to venues such as popular restaurants but also negative ratings, because they also wrote about those with a bad impression. Therefore, we can assume that the stars of non-obscure reviews are varied values. As a result, we believe that if distributions on stars of obscure and non-obscure are different and are similar to the above explanation, our classifier may classify reviews correctly.

Figure 3 shows the distributions of proportions of stars on obscure and non-obscure reviews. Here, blue bar shows the proportions of obscure reviews in each star value. Also, red bar shows the proportions of non-obscure reviews in each star value.

In Figure 3, the distributions of stars on obscure and non-obscure reviews are clearly different. The stars of obscure reviews are biased toward higher values. On the other hand, the stars of non-obscure reviews are evenly distributed. Therefore, Figure 3 shows that our classifier could classify reviews into obscure and non-obscure appropriately.

V. Analysis of obscure reviews and venues

In this section, we analyze obscure reviews and obscure reviews by using our classification methods. First, we discuss the obscure reviews and venues extracted by our method. Next, we discuss the categories of obscure venues. Finally, evaluate and discuss the differences in which each reviewer evaluates a venue as obscure or not.

A. Analysis of obscure review

This section describes and discusses obscure reviews extracted by using our proposed classifier.

In this experiment, we apply the classifier to all reviews. We used the document vectorization is SBERT and the classification algorithm is SVM, because this combination indicated the best performance in Section IV-D.

The number of reviews classified as obscure reviews is 312,151 (this is approximately 15% in all reviews). Table VIII shows some example of obscure reviews. Here, we replaced the name of the venue into "@" to anonymize it. In terms of review No. 1 of Table VIII, the reviewer wrote the location of the venue is negative but the food is positive. There are such texts in reviews classified as obscure reviews. The review No. 2 was written about a restaurant and the text contains the phrase "hidden gem". This phrase is a metaphorical expression for representing a place not very well known or unexpected find. Also, the review No. 3 and No. 4 contains the phrase representing obscure venues, but our obscure words in Table I does not include them. Therefore, their result shows that our classifier can find obscure reviews even if the texts do not directly contain the obscure words or phrases we have not prepared.

Also, we confirmed more obscure reviews manually. As a result, those reviews include many phrases of "I knew for the first time," "It was hard to access, but the service was good," and so. These phrases seem to be related to obscurity. Therefore, we believe that our method discovers venues that people have evaluated as obscure.

B. Classification results of obscure venue

This section describes and discusses the evaluation results of discovering an obscure venue using a classifier. In this experiment, we apply the classifier to all reviews of a venue and calculate the percentage of reviews classified as obscure.

The number of venues which were classified as the obscure venue is 10,915 (this is approximately 6% in all venues). Figure 4 shows the histogram of proportion of obscure reviews which were classified by our methods. This figure uses bins that are separated from 1.0 to 0.0 in 0.05 units. Here, for 1.0 and 0.5 in the figure, due to the small number of reviews included in the venue, the value is large. Venues without obscure reviews dominant in Figure 4. Also, most venues have a small percentage of obscure reviews. However, some venues have a high percentage of obscure reviews, and we regard such venues as obscure venues. Therefore, we believe that our approach can discover obscure venues.

TABLE VIII. Examples of reviews classified as obscure review.

No.	Text
1	Awesome place to eat! It may not look like much on the outside, but trust me...this place has some of the tastiest food in town.
2	This is a hidden gem. The decor is mixed but the food is excellent.
3	This is a great west side secret and I will be sure to refer the many people I encounter with in my position and let them know where I got my nails and toes done!
4	Located at the less well known spot of the @, the food court is less busy in comparison, thus it's never a hassle to find an empty seat.

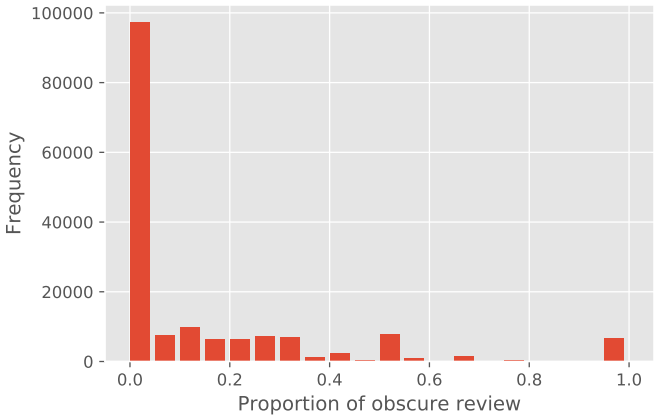

Figure 4. Histogram of proportion of obscure reviews in each venue.

C. Analysis of obscurity in each category

In this section, we analyze the obscure venues in each category. We denote the venue where the percentage of obscure reviews is 50% or more, according to the description in Section III-E, and find the proportion of venues classified as obscure within the same category.

We calculate the proportion of venues classified as obscure within a category. Here, the dataset from Yelp has 1,300 categories. Also, the venue in Yelp has at least one category. We used 62 categories whose number of reviews in a category is 1,000 or more.

We show the top 30 categories with the percentage of obscure venues in each category, as indicated in Figure 5. In this figure, the vertical axis shows the proportion of venues classified as obscure within the same category, and the horizontal axis shows the category names in Yelp. The highest percentage of obscure venues is for "Arts & Entertainment" at approximately 25%. This category has 3,886 venues in Yelp and 986 venues were classified as obscure venues. "Arts & Entertainment" has various subcategories in Yelp such as "Museums", "Stadiums & Arenas", and "Planetarium". However, these subcategories are not included in the ranking.

In Figure 5, the 2nd and 3rd categories are "Active Life" and "Shopping" at approximately 25%, respectively. "Active Life" has subcategories such as "Fitness & Instruction", "Baseball Fields", and "Parks". "Fitness & Instruction" is ranked at 18th. The obscure venue of this category occupies 25% of "Active Life". Also, "Shopping" has various various subcategories such as "Women's Clothing", "Fashion", and "Home & Garden". These subcategories are included in the top 30. Therefore, reviewers are likely to think of these subcategories as obscure venues.

In addition, according to Figure 5, the top categories with a high percentage of obscure venues contain many categories used in daily life such as "Shopping", "Education", and "Bakeries". In contrast, the subcategories of "Restaurants "such as "Steakhouses" , "Seafood", and "Breakfast & Brunch" where many people go to popular venues ranked the low. In these categories, popular venues are sometimes a type of sightseeing spot. However, as described in Table VIII, some venues were classified as obscure venues by our classifier. Therefore, we believe that such a result is correct as an analysis of obscure venues by categories.

D. Differences between venues evaluated as obscure for each reviewer

This section analyzes the differences among venues considered by reviewers as obscure.

Herein, we show the difficulty of providing a unique definition for obscure venues using our proposed method for obscure venue extraction. Using the classifier described in Section III-D, we classify whether a user review on a venue is obscure or not. Then, if the types of reviews on the venue are different, the venue that the user feels is obscure is different.

This research focused on cases in which two different reviewers posted similar reviews on two venue pairs. Two patterns of venues whose reviews refer to obscurity were considered, as shown in Figure 6. Pattern ① is a case in which two reviewers posted an obscure review and a non-obscure review to different venues. This pattern represents a case in which the reviewer felt that the referred venue was different. Pattern ② is a case in which the reviews posted by two different reviewers are the same for the referred venues. This pattern is one in which the venues the reviewers felt as obscure are the same. Therefore, if there is a certain number of reviews considered as pattern ①, it can be said that the venue perceived as obscure is different for each reviewer; the classification of obscure reviews reveals the contribution of the identification of obscure venues.

The procedure of this experiment is as follows. First, obscure venues to which two users posted similar reviews were extracted. During this experiment, 10,915 obscure venues that had obscure reviews were extracted, comprising more than 50% of all reviews. The classifier was then applied to the written reviews as described in Section IV. The numbers of the two patterns were calculated based on the classification results.

Table IX shows the experimental results. From Table IX, pattern ① comprised approximately 74% of the total. In other words, the combination of 74% of reviewers differs from the venue that was perceived as obscure. This result shows that the venues perceived as an obscure venue are not necessarily the same for all reviewers. Therefore, the approach of abstractly treating as obscure a review that includes an obscure word without criteria on the obscure venue used to extract the venue has the potential to be effective.

VI. Conclusion

In this research, we proposed a method for identifying obscure venues by extracting reviews that include descriptions

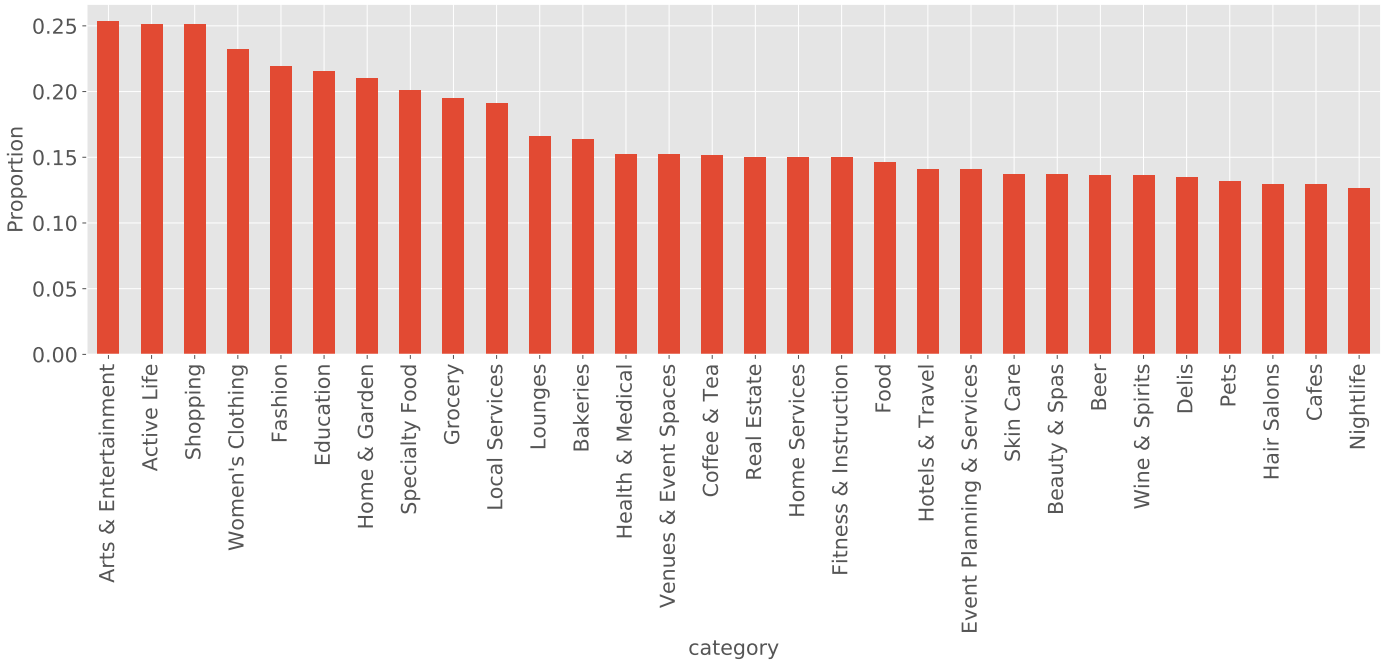

Figure 5. The top 30 categories with a high percentage of obscure venues in each category.

TABLE IX. PERCENTAGE OF DIFFERENCES IN REVIEWERS FEELING A VENUE AS BEING OBSCURE.

Pattern ①	17,206
Pattern ②	23,234
① / (① + ②)	0.74

regarding obscure posts on Yelp. Through reviews that include obscure words, a classifier was created to differentiate the reviews describing obscurity from those that do not, based on reviews in which the reviewers recognize the venues as being obscure. Evaluation results showed that the classifier is useful for extracting obscure reviews. Also, we discussed the differences of stars of obscure and non-obscure reviews to evaluate our method qualitatively. Furthermore, this research formulated and verified the hypothesis that venues perceived as obscure by reviewers are different. As a result, the venues perceived as being obscure are not necessarily the same for all reviewers, and our hypothesis is useful for discovering obscure venues.

Future studies will include a more detailed experiment and analyze obscure venues and the various categories present in each city. This paper is limited to analyzing obscure venues extracted using our proposed method in a qualitative manner. For a discovered venue, it is necessary to analyze whether it is obscure or not and to evaluate how useful or helpful the information is. For this purpose, we will conduct questionnaires by evaluators on the obscure venues by our proposed method. Further studies may apply our classifier to more various reviews such as another review site to discover obscure venues.

Also, there is necessary to examine the validity of obscure words. In this research, we used 13 obscure words, as Table shown in I. The experimental results represented that these obscure words are effective. However, we do not confirm that these words account for the majority of this meaning. The future work about obscure words investigates the validity of those words by questionnaires.

Also, the performance improvement of our obscure classifier is other future research. Although we used the LightGBM, SVM, and RF for this study, various methods for classification of texts such as graph convolutional network [42] and recurrent neural network [43] have been proposed. Also, because the number of reviews within obscure words is few, semi-supervised learning methods such as self-training [44] and label propagation [45] are suitable approaches for the situation. Those approaches might improve the obscure classifier, and we can extract more obscure reviews.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 16K00157, 16K16158 and 19K20418, and Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas Research on social big data.

REFERENCES

[1] M. Hirota, M. Endo, and I. Hiroshi, "Identifying obscure venues using classification of user reviews," in Proceedings of The Eleventh International Conference on Advances in Multimedia, ser. MMEIDA 2019, Mar 2019, pp. 7–12, ISBN:978-1-61208-697-2, URL:http://ns2.thinkmind.org/index.php?view=article&articleid=mmedia_2019_1_20_58003.

[2] "Yelp," URL: https://www.yelp.com/ [accessed: 2019-02-27].

[3] "Expedia," URL: https://www.expedia.com/ [accessed: 2019-02-27].

[4] "Tripadvisor," URL: https://www.tripadvisor.com/ [accessed: 2019-02-27].

[5] D. Ukpabi, S. Olaleye, E. Mogaji, and H. Karjaluoto, "Insights into online reviews of hotel service attributes: A cross-national study of selected countries in africa," in Information and Communication Technologies in Tourism 2018, B. Stangl and J. Pesonen, Eds. Cham: Springer International Publishing, 2018, pp. 243–256.

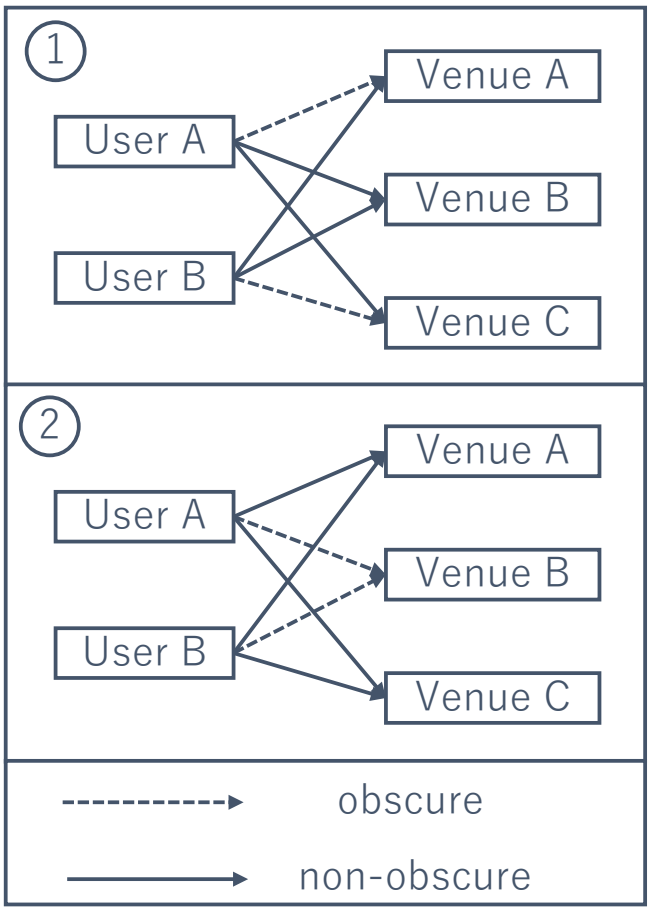

Figure 6. Pattern in which two reviewers evaluate venues as obscure.

[6] V. Browning, K. K. F. So, and B. Sparks, "The influence of online reviews on consumers' attributions of service quality and control for service standards in hotels," Journal of Travel & Tourism Marketing, vol. 30, no. 1-2, 2013, pp. 23–40.

[7] C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa, "Anaba: An obscure sightseeing spots discovering system," in 2014 IEEE International Conference on Multimedia and Expo, 2014, pp. 1–6.

[8] D. Kitayama, "Extraction method for anaba spots based on name recognition and user's evaluation," in Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, ser. iiWAS '16. ACM, 2016, pp. 12–15.

[9] S. Katayama, M. Obuchi, T. Okoshi, and J. Nakazawa, "Providing information of hidden spot for tourists to increase tourism satisfaction," in Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, ser. UbiComp '18. ACM, 2018, pp. 377–380.

[10] Z. Liu and S. Park, "What makes a useful online review? implication for travel product websites," Tourism Management, vol. 47, 2015, pp. 140 – 151.

[11] M. Toyoshima, M. Hirota, D. Kato, T. Araki, and H. Ishikawa, "Where is the memorable travel destinations?" in Social Informatics. Cham: Springer International Publishing, 2018, pp. 291–298.

[12] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation," Tourism Management, vol. 59, 2017, pp. 467 – 483.

[13] C. Vo, D. Duong, D. Nguyen, and T. Cao, "From helpfulness prediction to helpful review retrieval for online product reviews," in Proceedings of the Ninth International Symposium on Information and Communication Technology, ser. SoICT 2018. ACM, 2018, pp. 38–45.

[14] Z. Wang, B. S. Balasubramani, and I. F. Cruz, "Predictive analytics using text classification for restaurant inspections," in Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, ser. UrbanGIS'17. ACM, 2017, pp. 14:1–14:4.

[15] W. He, X. Tian, R. Tao, W. Zhang, G. Yan, and V. Akula, "Application of social media analytics: a case of analyzing online hotel reviews," Online Information Review, 2017.

[16] R. Dong, M. Schaal, M. P. O'Mahony, and B. Smyth, "Topic extraction from online reviews for classification and recommendation," in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI '13. AAAI Press, 2013, pp. 1310–1316. [Online]. Available: http://dl.acm.org/citation.cfm?id=2540128.2540317

[17] M. Afzaal, M. Usman, A. C. M. Fong, S. Fong, and Y. Zhuang, "Fuzzy aspect based opinion classification system for mining tourist reviews," Advances in Fuzzy Systems, 2016, pp. 1–14.

[18] T. S. Bang and V. Sornlertlamvanich, "Sentiment classification for hotel booking review based on sentence dependency structure and sub-opinion analysis," IEICE Transactions on Information and Systems, vol. E101.D, no. 4, 2018, pp. 909–916.

[19] P. Phillips, S. Barnes, K. Zigan, and R. Schegg, "Understanding the impact of online reviews on hotel performance: An empirical analysis," Journal of Travel Research, vol. 56, no. 2, 2017, pp. 235–249.

[20] S. Karimi and F. Wang, "Online review helpfulness: Impact of reviewer profile image," Decision Support Systems, vol. 96, 2017, pp. 39 – 48.

[21] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment analysis in tourism: Capitalizing on big data," Journal of Travel Research, vol. 58, no. 2, 2019, pp. 175–191.

[22] E. van der Zee, D. Bertocchi, and D. Vanneste, "Distribution of tourists within urban heritage destinations: a hot spot/cold spot analysis of tripadvisor data as support for destination management," Current Issues in Tourism, vol. 23, no. 2, 2020, pp. 175–196.

[23] L. Jhih-Yu, W. Shu-Mei, H. Masaharu, A. Tetsuya, and I. Hiroshi, "Less-known tourist attraction analysis using clustering geo-tagged photographs via x-means," International Journal on Advances in Systems and Measurements, vol. 12, no. 3&4, 2019, pp. 215–224.

[24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, 2017, pp. 135–146.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[28] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2019.

[29] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018.

[30] A. Adhikari, A. Ram, R. Tang, and J. Lin, "Docbert: Bert for document classification," 2019.

[31] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[32] "langdetect," URL: https://pypi.org/project/langdetect/ [accessed: 2020-02-20].

[33] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language

Processing (EMNLP). Association for Computational Linguistics, 2015.

[34] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, 2018, pp. 1112–1122. [Online]. Available: http://aclweb.org/anthology/N18-1101

[35] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, 1995, pp. 273–297.

[36] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, 2001, pp. 5–32.

[37] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3146–3154.

[38] "Yelp dataset challenge (round 13)," URL: https://www.yelp.com/dataset/challenge [accessed: 2020-02-20].

[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," 2019.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[41] "Lightgbm, light gradient boosting machine," URL: https://github.com/microsoft/LightGBM [accessed: 2019-02-27].

[42] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 7370–7377.

[43] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," 2016.

[44] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and lda topic models," Expert Systems with Applications, vol. 80, 2017, pp. 83–93.

[45] Z.-W. Zhang, X.-Y. Jing, and T.-J. Wang, "Label propagation based semi-supervised learning for software defect prediction," Automated Software Engineering, vol. 24, no. 1, 2017, pp. 47–69.