2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Opinion Classifier Transfer Learning From Review Data

Jin Ozeki
Graduate School of Advanced Mathematical Sciences, Meiji University
Tokyo, Japan
cs233005@meiji.ac.jp

Yoshitaka Sakurai
School of Interdisciplinary Mathematical Sciences, Meiji University
Tokyo Japan
sakuraiy@meiji.ac.jp

Yuna Terada
TSP Inc.
Tokyo, Japan
yuna.tera.555@gmail.com

Abstract*— Companies use users' opinions to improve their products and marketing activities. In recent years, the development of Internet technology has made it possible to extract users' opinions from text on the Web. There are many ways for users to post their opinions on the Internet, and Twitter is considered to be a platform that allows users to easily tweet their opinions. However, manually extracting opinions from Twitter is time-consuming, costly, and labor-intensive due to the relatively low percentage of opinions. Therefore, some companies aim to efficiently extract user opinions from Twitter using machine learning. However, the attempt to create a dataset for building a machine learning system produced an unbalanced dataset that does not extract opinions with sufficient accuracy because the proportion of views on Twitter is small. There are solutions to this problem of insufficient teacher data, such as utilizing knowledge from other domains through transfer learning. Although transfer learning is sometimes used to solve such problems, accuracy cannot be improved if the knowledge domains are far apart before and after the transfer. Therefore, we proposed a new method called OTR, which stands for Opinion classifier Transferred from Review data. OTR transfers knowledge of review submissions that are considered to be close in domain to opinion extraction. However, since the phrasing of review sentences and that of Social Networking Service (SNS) such as Twitter are different, there is a possibility that sufficient knowledge transfer cannot be achieved. In order to address this problem, we proposed an Opinion classifier Transferred from Review data with Pseudo-labels (OTR-P), a method that brings the domains of the source and target tasks closer. Here, the target task discriminated opinions regarding leisure facilities, and the source task estimated review ratings using Rakuten travel review data. And while performing these tasks, we attempted to bring the domains closer by attaching pseudo-labels to the tweet data. This approach improved accuracy compared to the conventional method of shifting Bidirectional Encoder Representations from Transformers (BERT).

Keywords*— *Transfer Learning, Pseudo-labels, Opinion Mining, Domain Adaptation.

I. INTRODUCTION

Many companies improve their services and products based on the opinions of users (consumers) collected via conventional methods such as questionnaires. However, user contributions to Social Networking Service (SNS) and other consumer-generated media have become increasingly important in recent years due to the Internet's spread.

However, the proportion of opinions useful for improving services on Twitter is small, and manual analysis requires considerable effort. Therefore, creating an opinion extraction model using Twitter posts as teacher data is conceivable to extract opinions automatically.

This type of research, which collects opinions from texts on the Internet, is known as opinion mining.

In this study, we considered a machine learning method to efficiently extract user opinions from SNS. However, while attempting to create a dataset for constructing a machine learning system, we obtained an unbalanced dataset because the proportion of opinions on Twitter was small. This makes it challenging to construct a highly accurate opinion-extraction model.

The lack of teacher data is a well-known challenge in machine learning, and one solution is transfer learning. Transfer learning is a technique that enables the construction of highly accurate models, even with a small amount of target-teacher data, by transferring knowledge from another domain (source task) to the target task (target task).

In this study, we propose a method for constructing an opinion classification model from SNS using machine learning and transition learning to make the model highly accurate while using a small amount of supervised data.

Opinion mining is defined as the extraction of opinions from the Web. It analyzes people's opinions, appraisals, attitudes, and emotions toward organizations, entities, persons, issues, actions, topics, and attributes [1].

Sohrabi et al. also summarized the key points of sector-wise emphasis on opinion mining, and while extracting opinions from the company's point of view, those that helped improve the product were considered significant [2].

Therefore, based on the above definitions and perspectives, in this study, we consider opinion extraction to be "collecting opinions useful for service improvement."

Twitter, a major SNS, was used as the target medium in this study, and opinion extraction was conducted by determining whether a Twitter post is an opinion or not.

In transfer learning, it is generally believed that the closer the target and source domains are, the better the accuracy.

Review evaluation estimation can be considered a task similar to opinion discrimination, which is the target task of this study. Therefore, we examined the effects of transfer learning using review data as the source. In this study, we developed a model for extracting opinion from tweets related to leisure facilities. Our target task involved using Rakuten travel review data, which are widely used for research purposes. Through our investigation, the effectiveness of Rakuten travel data as a source was verified.

Rakuten Travel reviews are texts related to opinions and are considered a domain close to the objective of discriminating opinions regarding leisure facilities. However, they have different characteristics from text because they are not tweets, which may cause accuracy degradation. Therefore, we propose a method to inherit knowledge from review data using pseudo-labels while maintaining the characteristics of tweets and testing its effectiveness.

In this study, we propose a method for improving the accuracy of Twitter posts using transfer learning to classify opinions, and we verify the effectiveness of this method. We clarified the usefulness of the proposed method for discriminating opinions in review data, particularly Rakuten travel reviews, and the effect of improving accuracy by creating training data in the form of tweets using pseudo-labels.

II. RELATED RESEARCH

A. Opinion Mining

Opinion extraction is performed as a classification task to determine whether a text contains an opinion. Text classification methods can be broadly classified into dictionary-and machine-learning-based methods.

A dictionary-based method performs classification according to rules created in advance by humans. For example, words such as good, like, and beautiful are considered positive, whereas words such as bad, dislike, and dirty are considered negative.

A study of dictionary-based opinion extraction by Tateishi et al. [3] created a dictionary by classifying expressions indicating human sensations and feelings, such as "good" and "like," and expressions indicating properties and characteristics of things, such as "fast" and "small," into positive and negative, and constructed a system to extract opinions by using pattern matching rules.

Research on emotion analysis has been conducted more widely for Twitter text classification than opinion extraction. Pandarachalil et al. proposed a dictionary-based study of Twitter sentiment analysis [4], in which they created three emotion dictionaries for Twitter accents and analyzed these emotions.

Thus, dictionary-based methods are effective if suitable dictionaries can be created for each domain and language; however, creating such dictionaries is expensive [2].

It should be noted that the dictionary base cannot handle tasks in complex contexts, such as Twitter [5].

Therefore, in this study, we utilized machine learning to perform opinion discrimination.

Machine-learning-based methods can be broadly divided into supervised, unsupervised, and semi-supervised learning methods.

In supervised learning, data for which correct answers have already been labeled are utilized.

For example, Enbrahimi et al. proposed a neural network-based classifier to identify predatory conversations in chat logs automatically and demonstrated its functionality [6].

However, it can be difficult to prepare a sufficient amount of supervised data for supervised learning. This is the case when annotation work is expensive and time-consuming or when the annotator must have expert knowledge. Transition learning, or semi-supervised learning, can be considered a solution to the problem of a small amount of teacher data.

Transfer learning is a technique to improve learning efficiency by transferring learned knowledge from one domain to another.

This technique was used in Bidirectional Encoder Representations from Transformers (BERT) [7]. BERT is a pre-trained model presented by Google in 2018 that uses a large text corpus to perform tasks such as sentence classification with high accuracy, using the pre-trained model as the source model and transferring knowledge from a small amount of labeled data to the target.

Thus, transition learning enables the construction of highly accurate models even with a small amount of target teacher data.

This study created a source model using BERT and reviewed the data. And opinion discrimination model was created by training a small amount of leisure facility data as targets.

In addition, semi-supervised learning may be used as a solution for small amounts of teacher data. Semi-supervised learning is a method for increasing the amount of teacher data by creating a model using annotated data and labeling the unannotated data with the model. Semi-supervised learning is often used because it is easy to collect unlabeled data, particularly in web data and Twitter environments.

Semi-supervised learning for Twitter sentiment analysis was used in studies by Silva et al. and Hong et al. [8][9].

Da Silva et al. proposed a self-learning algorithm using semi-supervised learning because of the difficulty in creating labeled datasets on Twitter.

Hong et al. used a self-training model for Twitter classification. This study demonstrated that self-training improves Twitter classification accuracy.

Thus, semi-supervised learning is a solution for small amounts of supervised data.

The above semi-supervised learning is self-learning in the same domain as the target; however, in this study, to effectively transfer the review classification knowledge, a review classification model was first built using the annotated data, as in semi-supervised learning, and thereafter, the model was used to annotate the unannotated data. Thus, we annotated the Twitter text so the domain was closer to the target task.

III. Proposed Method

In this study, we propose an **O**pinion classifier **T**ransferred from **R**eview data (**OTR**), a transfer learning method for review evaluation, and an **O**pinion classifier **T**ransferred from **R**eview data with **P**seudo-labels (**OTR-P**), a met technique that uses pseudo-labels.

Reviews reflect user opinions because users experience and evaluate services and products. In addition to examinations, word of mouth reflects our views and opinions; however, word of mouth contains vague information such as rumors and hearsay. In contrast, reviews are written based on users' experiences and evaluated numerically; therefore, they can be easily used as teacher data for machine learning systems. Therefore, in this study, we construct a review classification model using review ratings and transfer this knowledge to an opinion discrimination task.

Additionally, we created a source model pseudo-labeled by the review classification model to further approximate the domain.

In the following sections, we present the modeling procedure for OTR and OTR-R.

A. *OTR: Opinion classifier Transferred from Review data*

In this method, a review classification model is first built. Subsequently, opinion discrimination is performed via transfer learning. Although it is possible to construct a review classification model by learning from the review dataset alone, transferring knowledge using a language model, which is the basic design of this method, is generally more accurate. Any language model can be used if its output is classified. In the case of the production of the review classification model, because a review dataset is often rated on a five-point scale, it can be classified into five categories or converted into a binary classification.

The procedure for the method OTR shown in Fig. 1 is as follows:

Step 1: Review Classification Model building

Transition Learning from Language Model to Review Data

- Data used: review data set
- Input: text (review sentences)
- Output: classification, number of classes = two (positive-negative), three, or five

Step 2: Opinion discrimination model construction

Transfer learning from the review classification model

Data: Opinion classification dataset

- Input: text (SNS postings)
- Output: classification, two classes (opinion and non-opinion)

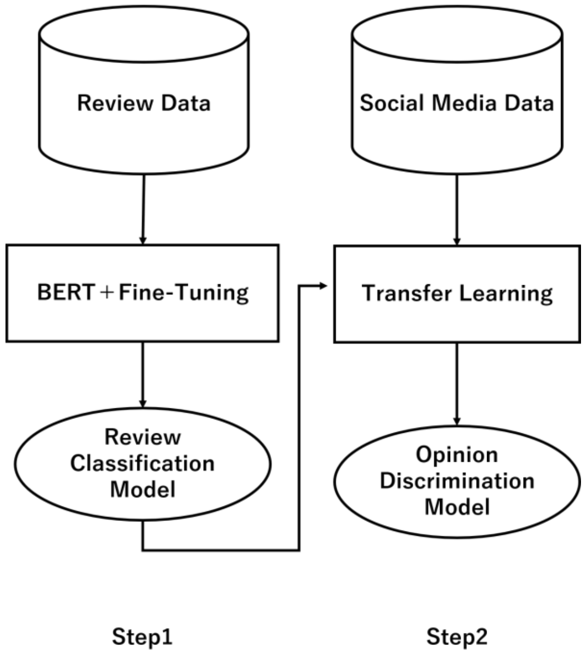

Fig. 1. Opinion classifier transferred from review data procedure.

B. *OTR-R: Opinion classifier Transferred from Review data with Pseudo-label*

The review data were posted on review sites, whereas the opinion classifier dataset was posted on Twitter and other sites. Because the media are different, the text's length, assumed information, and style differ. Therefore, in the opinion classifier transferred from review data with pseudo-labels (OTR-P), a pseudo-label is attached to the treader in a format similar to the target in the review classification model. A pseudo-label data classification model is constructed, bringing the domains closer via transfer learning. The domain was approached by building a pseudo-label data classification model and using transfer learning.

The procedure for the method OTR shown in Fig. 2 is as follows:

Step 1: Review Classification Model building

(same as Step 1 in Fig. 1).

Step 2: Creation of pseudo-label data

Label data from the same media as the target task using the classification results in the review classification model.

- Target data: posting data from the same media as the opinion discrimination dataset.

Step 3: Same-mediatization review classification model construction

- Data used: pseudo-label data created in Step 2, review data set
- Input: text (review sentences, SNS posts)

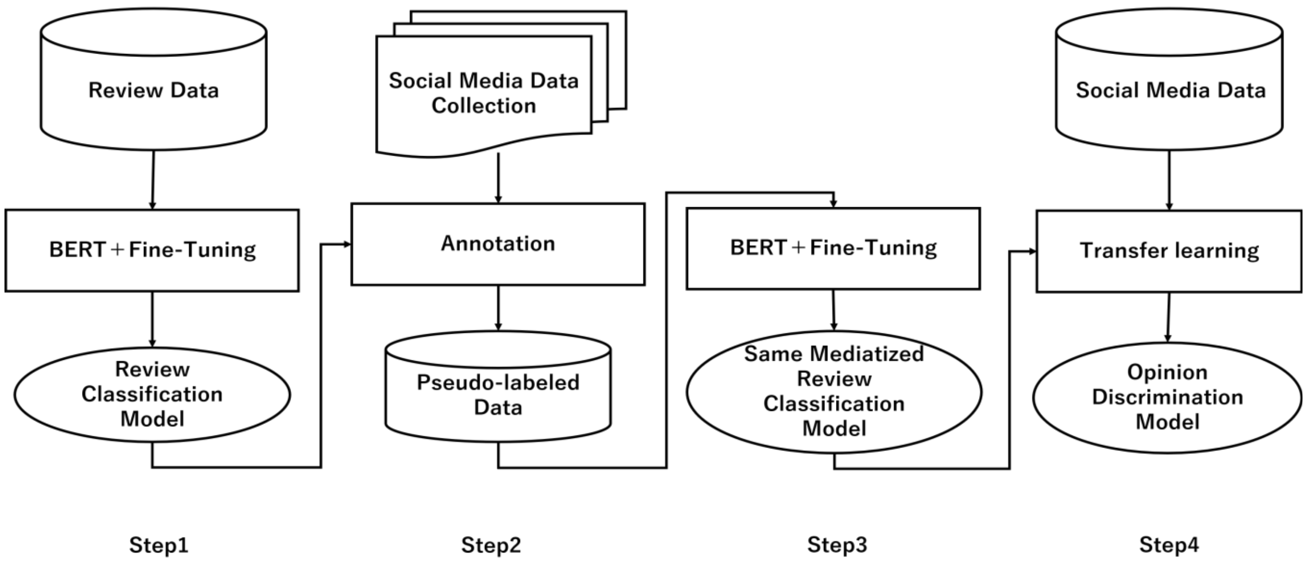

Fig. 2. Opinion classifier transferred from review data with pseudo-label procedure.

- Output: classification, number of classes = two (positive-negative), three, or five

Step 4: Opinion Discrimination Model building

Transition Learning from Pseudo-labeled Data Classification Model

Data used: opinion-discriminant datasets

- Input: text (SNS posts)
- Output: classification, two classes (opinion, non-opinion)

In the construction of a homogenized review classification model, not only pseudo-label data but also datasets for the construction of the review classification model can be used, and by changing the utilization ratio of each, the influence of the original review dataset may vary. Therefore, adjustments are necessary, depending on the dataset used.

IV. EXPERIMENT

A. Experiment

In this experiment, we improved the accuracy of opinion discrimination by transferring knowledge from the review data.

Moreover, we compared OTR with OTR-P to verify the effect of bringing the domains closer using pseudo-labels. Although OTR-P can increase the number of data points using the number of unlabeled data points in the same media as opinion discrimination, in this study, the number of training data was unified to 4320. The OTR setup used in the experiment is described in the following sections.

1) Used Review Dataset: Rakuten travel review data were used in the study [10]. Rakuten Travel review data consisted of approximately 29,000 facility data points and approximately 6.59 million review data points (posting date and time, plans, purpose of use, user-submitted text, overall rating, etc.); we created the model via the following process:

Step 1: Extraction of user-submitted text and overall values from Rakuten travel review data.

Step 2: Postings with an overall rating of one or two were considered negative, and those with a rating of four or five were considered positive.

Step 3: Random sampling was performed to ensure an equal number of data points for each class, split 8:1:1 for training, validation, and testing.

2) Pre-trained language models were used: The Japanese pre-trained BERT model published by Kyoto University was used as the pre-trained model [11]. This model was pre-trained using the Japanese Wikipedia.

3) Pretreatment method: The review text was word-segmented and converted to IDs for input into the pre-trained model. Juman++ was used for the morphological analysis, which considers the semantic naturalness of word sequences using a recurrent neural network language model (RNNLM) [12]. In addition, because the text input to the BERT must be of a fixed length, all sentences in this study were converted into 384 words.

4) Setting up transfer learning

In the transfer learning method, an output layer was added to the last layer of the pre-trained model to perform review submission classification, and fine-tuning was performed using the training data. The parameters were set as follows.

- Batch size: 32
- Optimizer: Adam
- Learning rate: 2e-5
- Loss function: CrossEntropyLoss

To prevent overlearning, early stopping was used to discontinue learning if the data loss for validation did not decrease over five epochs in a row.

5) Opinion-classification data set

The target opinion identification task uses a dataset constructed by Nozaki et al. for opinion extraction [5]. Below is an overview of the opinion definition stations.

The dataset for opinion extraction aims to collect "opinions useful for service improvement" and includes tweets that express "requests" and "criticisms/feelings." However, some "requests" and "criticisms/feelings" cannot be applied to marketing. For example, tweets that do not include a reason or cause of the emotion, such as "Disney is fun," are not beneficial from a marketing perspective. Therefore, Nozaki et al. limited the expressions to those that indicated "why such feelings or criticisms were expressed" as opinions

Thus, we used a dataset annotated with "requests" and "criticisms and feelings" expressed as opinions. Subsequently, we targeted a binary classification task, classifying whether a post is an opinion.

6) Setting of pseudo-labels

Tweets were collected using target leisure facilities as keywords and annotated using a review classification model. The procedure was as follows:

6-1: Data collection in the same format as the target

Because the aim is to identify user opinions on Twitter, tweets containing the keyword "Disney" were collected using the Twitter API. The number of collected Tweets was 10930.

6-2: Data preprocessing

Relevant sections containing pictograms and URLs were deleted and regularized to be fed as inputs to the classification task.

6-3: Annotation with the review classification model

The review classification model classified tweet data as negative or positive and annotated the prediction results. The annotation results are listed in Table 1.

TABLE I. ANNOTATION RESULTS

Class	Number of Tweets
Negative	8216
Positive	2714

6-4: Under sampling

To eliminate bias in the number of data points, we sampled the pseudo-labeled tweet data to obtain a total dataset of 2,700 negatives and 2,700 positives, for a total of 5,400 datasets. And 80% of this data, or 4,320 data, will be used for training.

7) Libraries used in the experiment

The study was programmed using Python version 3.8.10. The main libraries used in the experiments are as follows:

- mojimoji 0.012
- numpy 1.22.4
- pandas 1.5.3
- pip 23.1.2
- torch 2.0.1+cu118
- torchtext 0.15.2
- transformers 4.31.0
- pyknp 0.6.1

8) Number of data points in the source model

The number of data points for the model used in the Experiment is listed in Table 2. In the Experiment, only the pseudo-label data were used as training data for the OTR-P homogenized review classification model to verify the effect of using pseudo-labels, and the number of data points was the same as the review data points.

TABLE II. NUMBER OF DATA POINTS IN SOURCE MODEL (EXPERIMENT)

	Review	*Pseudo-label*
Base	0	0
OTR	4320	0
OTR-P	0	4320

B. Experiment Results

The results of Experiment are listed in Table 3.

The base method with the transfer learning of BERT produced relatively low values for all indicators, whereas OTR-P with the proposed method produced relatively high values.

TABLE III. EXPERIMENT RESULTS

	BERT	*OTR*	*OTR-P*
Accuracy	0.67	**0.71**	0.70
Precision	0.65	**0.68**	0.65
Recall	0.72	0.79	**0.85**
F1-score	0.69	0.73	**0.74**

C. Experiment Discussion

Comparing the base method and OTR, OTR had higher values. In particular, OTR had higher values for the F1-score, which is considered necessary for classification problems. This suggests that the accuracy of opinion discrimination is improved by transferring knowledge from review data.

The accuracy of OTR was higher than that of OTR-P. In addition, the precision value decreased and the recall value increased. This is because the OTR-P judges more tweets as opinions. Furthermore, the F1-score of the synergistic average of accuracy and precision improved. This suggests that OTR-P could collect user opinions without omissions and improve its performance over OTR.

Thus, pseudo-labels allow us to bring the domains closer to each other.

V. CONCLUSION

In this study, we tested OTR, a method for transferring the knowledge of word-of-mouth ratings to discriminate opinions.

By using review-posting data, we constructed an opinion discrimination model with high accuracy when transfer learning was performed. In addition, annotating tweet data with pseudo-labels allows us to bring the domains closer together.

In addition, owing to the number of samples used in this study, annotations in the review classification model were labeled with positive and negative values. Furthermore, we consider it necessary to set the output of the review classification to five levels when extracting detailed opinions. Other areas of the study need to be validated to determine whether they are effective in places other than review facilities.

VI. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 20K11960. We are grateful to "the Rakuten Group, Inc. for providing us with the "Rakuten Dataset" (https://rit.rakuten.com/data_release/) via the IDR Dataset Service of the National Institute of Informatics.

REFERENCES

[1]. B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies Morgan & Claypool Publishers, 1-167, 2012.

[2]. F. Hemmatian, and M. K. Sohrabi, "A survey on classification techniques for opinion mining," Artif. Intell. Rev., vol. 52, pp. 1495–1545, 2017.

[3]. K. Tateishi, Y. Ishiguro, and T. Fukushima, "A Reputation Search Engine from the Internet," Journal of Japanese Society for Artificial Intelligence, vol.19, no.3, pp. 317–323, 2004, (in Japanese).

[4]. R. Pandarachalil, S. Sendhilkumar, and G. S. Mahalakshmi, "Twitter sentiment analysis for large-scale data: an unsupervised approach," Cogn. Comput., vol. 7, pp. 254–262, 2015.

[5]. Y. Nozaki and Y. Sakurai, "Dataset Creation Method for Constructing Opinion Mining Model from Tweets," Transactions on Mathematical Modeling and its Applications (TOM), Vol.13, No.2, pp.69-83, 2020, (in Japanese).

[6]. M. C. Y. Ebrahimi, and O. S. Ormandjieva, "Detecting predatory conversations in social media by deep convolutional neural networks," Digit Investing, vol. 18, pp. 33–49, 2016.

[7]. J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT) , Vol.1, pp.4171-4186 2019.

[8]. N. F. F. da Silva, L. F. S. Coletta, E. R. Hruschka, and E. R. Hruschka Jr., "Using unsupervised information to improve semi-supervised tweet sentiment classification," Inf. Sci., vol. 355–356, pp. 348-365, 2016.

[9]. S. Hong, J. Lee, J.-H. Lee, "Competitive self-training technique for sentiment analysis in mass social media," 15th international symposium on soft computing and intelligent systems, Kitakyusyu, Japan, pp. 9–12, 2014.

[10]. Rakuten GroupInc, Rakuten Travel data. Informatics Research Data Repository, National Institute of Informatics.(dataset). https://doi.org/10.32130/idr.2.2, 2020.

[11]. Kyoto University. "Pretrained japanese BERT model", Apache License, Version 2.0. : https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

[12]. H. Morita, D. Kawahara, and S. Kurohashi, "Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model", Proceedings of the 2015Conference on Empirical Methods in Natural Language Processing(EMNLP), pp.2292-2297, 2015.