DE GRUYTER
OPEN

Research Paper

Sentiment Analysis of Japanese Tourism Online Reviews

Chuanming Yu[1], Xingyu Zhu[1], Bolin Feng[1], Lin Cai[1], Lu An[2†]

[1]School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China
[2]School of Information Management, Wuhan University, Wuhan 430072, China

Citation: Chuanming Yu, Xingyu Zhu, Bolin Feng, Lin Cai, Lu An (2019). Sentiment analysis of Japanese tourism online reviews.
Vol. 4 No. 1, 2019
pp 89–113
DOI: 10.2478/jdis-2019-0005
Received: Oct. 15, 2018
Revised: Dec. 25, 2018
Accepted: Jan. 8, 2019

Abstract

Purpose: Online reviews on tourism attractions provide important references for potential tourists to choose tourism spots. The main goal of this study is conducting sentiment analysis to facilitate users comprehending the large scale of the reviews, based on the comments about Chinese attractions from Japanese tourism website 4Travel.

Design/methodology/approach: Different statistics- and rule-based methods are used to analyze the sentiment of the reviews. Three groups of novel statistics-based methods combining feature selection functions and the traditional term frequency-inverse document frequency (TF-IDF) method are proposed. We also make seven groups of different rules-based methods. The macro-average and micro-average values for the best classification results of the methods are calculated respectively and the performance of the methods are shown.

Findings: We compare the statistics-based and rule-based methods separately and compare the overall performance of the two method. According to the results, it is concluded that the combination of feature selection functions and weightings can strongly improve the overall performance. The emotional vocabulary in the field of tourism (EVT), kaomojis, negative and transitional words can notably improve the performance in all of three categories. The rule-based methods outperform the statistics-based ones with a narrow advantage.

Research limitation: Two limitations can be addressed: 1) the empirical studies to verify the validity of the proposed methods are only conducted on Japanese languages; and 2) the deep learning technology is not been incorporated in the methods.

Practical implications: The results help to elucidate the intrinsic characteristics of the Japanese language and the influence on sentiment analysis. These findings also provide practical usage guidelines within the field of sentiment analysis of Japanese online tourism reviews.

Originality/value: Our research is of practicability. Currently, there are no studies that focus on the sentiment analysis of Japanese reviews about Chinese attractions.

† Corresponding author: Lu An (E-mail: anlu2009@whu.edu.cn).

Research Paper

Keywords Sentiment analysis; Japanese reviews; Rule-based methods; Statistics-based methods; Tourism reviews

1 Introduction

In recent years, tourism, as an activity enriching life and widening sight, has not only promoted rapid economic development but also facilitated communication between various regions. Inbound tourism for a nation brings considerable foreign-currency revenue and is an important way to improve the national image. With the rapid development of the Internet, the advent of "Internet + travel" patterns has improved the efficiency of the traditional tourism industry. In 2017, in terms of Chinese outbound tourism report, the total number of visitors reached 131 million, 70% of which chose independent travel (Chinese Outbound Tourism Development Annual Report, 2018). In 2020, according to a forecast of industry insiders, the number of visitors will reach 250 million (Zhang, 2015). Compared with flourishing outbound tourism, the growth rate of inbound tourism for China is very limited. According to the statistics of the National Tourism Administration of China, the number of foreign visitors reached 4.4567 million in the fourth quarter of 2015, and the year-and-year growth was only 0.92% (NTASMD, 2015). In addition, there is a trade deficit between inbound and outbound tourism with respect to the neighboring countries. For example, the number of Chinese visitors to Japan accounts for 11% of outbound tourism, whereas the number of Japanese visitors accounts for just 5% of the total number of inbound tourists in China.

The combination of tourism and the Internet allows people to obtain more detailed travel information. Comments regarding attractions are becoming an increasingly important basis for the selection of tourist attractions. Mining the reviews on travel websites can help potential visitors better understand tourist attractions so that they can choose their favorite scenic spots and avoid and reduce trouble throughout the tour.

For the reasons mentioned above, there is a need to study Japanese reviews of Chinese attractions to improve the service of tourist attractions. In the scientific communities worldwide, a growing number of studies have focused on sentiment analysis of online reviews. There is a great need for new tools and algorithms which can automatically, efficiently and robustly process the large amounts of user-generated content that emerge daily. Most prior research on sentiment analysis has been carried out on English and other European languages (Contreras et al., 2018; Fernández & Sebastiani, 2016; Impana & Kallimani, 2017; Severyn et al., 2016; Soni, 2017; Zin et al., 2018). To date, there have been few studies concentrating on Japanese reviews. The number of studies on Japanese online reviews is considered disproportionate with the fact that Japan is the third largest economic body.

In this study, we focus on 4travel[①]. It is a platform that hosts a great variety of reviews uploaded by Japanese travelers. We conduct sentiment analysis of Japanese reviews on Chinese attractions utilizing statistics- and rule-based methods. Three groups of novel methods combining feature selection functions and the traditional term frequency-inverse document frequency (TF-IDF) method are proposed. Additional sentiment words and emoticons are chosen to complement the original emotional dictionary and an extensive comparative study is performed.

Compared with existing studies, the main contributions and the innovation of our research are as follows.

1) The combination of feature selection functions and the traditional TF-IDF method solves the problem of traditional TF-IDF being unable to fully represent the connection between features and categories. The influence of different feature selection functions, feature weightings and kernel functions on the sentiment analysis of Japanese online tourism reviews is explored. As will be experimentally shown in Section 4.1, the proposed combined methods outperform the traditional TF-IDF method and the SVM method with word embeddings in terms of macro and micro-average values. Also, in the Section 4.5, we will find the proposed methods perform better than all of the other popular methods for sentiment analysis, such as the methods based on neural networks.
2) Additional sentiment words and emoticons, including the emotional vocabulary in the field of tourism (EVT), kaomojis (emoticons), related emotional word (REWs) and terms with an implicit emotional tendency (TIETs), are used to complement the traditional Japanese *Kanjo Hyogen Jiten* (emotional expression dictionary). As will be experimentally shown in Sections 4.2 and Section 4.5, the EVT, kaomojis, negative and transitional words can improve the macro- and micro-average values considerably. The results help elucidate the intrinsic characteristics of the Japanese language and the influence on sentiment analysis.
3) It is proven that Mutual Information (MI) owns the highest ability of representation, even with a small number of features, whereas CHI holds the strongest noise-tolerant ability. The Laplace kernel obtains the highest priority of recommendation. Comparisons between the statistics- and rule- methods help to ease the controversy between the two types of methods and strengthen the evidence that the rule-based ones may achieve higher performance if the determination rules are well designed. These findings provide practical usage guidelines within the field of sentiment analysis of Japanese online tourism reviews.

[①] http://www.4travel.jp

The remainder of the paper is structured as follows. In Section 2, a comprehensive review is conducted on the statistics- and rule-based methods. In Section 3, the experimental setup is described, elaborating upon the features of the datasets, the evaluation metrics and the design of the comparative study between the baselines and proposed statistics- and rule-based methods. The experimental results of the proposed statistics- and rule-based methods are compared with those of other methods and discussed in Section 4. The study's significant findings and conclusions are presented in Section 5.

2 Related Works

At present, research on sentiment analysis has made some progress. Related methods can be mainly divided into statistics-based approaches and rule-based approaches. For statistics-based approaches, the key issue is feature selection and weighting in reviews; for rule-based ones, the focus is on the construction of emotional dictionaries and determination of rules.

2.1 Statistics-based approaches

The statistical approaches determine the sentiment of a document based on extracted sentimental features and the machine learning approaches. Palakvangsa-Na-Ayudhya., Sriarunrungreung, Thongprasan, and Porcharoen (2011) constructed a system for the tourism business to classify comments gathered from available travel social network websites into predefined aspects and further analyze comments into positive and negative sentiments. A manually created list of words is used to map each word obtained from the previous step into these predefined aspects. Also, a manually created list is used to map a given word to the polarity. The accuracy of the system is above 85% and more than 90% of the users give the overall rating equal or above the rate 8. However, in this system, two manually created lists are required so that it may cause inaccurate results. Ma and Deng (2013) proposed an approach of weighting features in short texts to resolve the problem of poor classification performance caused by an uneven distribution of classes in short texts. The approach improves the overall performance to a certain extent but does not obviously improve the precision and recall values in small categories. Xiao et al. (2015) argued that traditional Chi-square (CHI) is easy to give higher evaluation to some unimportant low-frequency words, especially in the case of the uneven distribution of the classes. The reason is, the method only considers whether the feature appears in this case, and thus ignores other useful information in the feature, which causes partiality to the low frequency words, exaggerates the role of the low frequency words, and easily introduces noise words in the process of feature selection. Thus, they set a threshold to remove low-frequency terms and calculate the CHI of the remaining terms. The new approach achieves higher performance than the original CHI while the promotion is not significant regarding small categories. Abd-Elhamid, Elzanfaly, and Eldin (2017) used the Part Of Speech (POS) tagging feature to execute automatically the extraction and weighting of sentiment features from a set of annotated reviews. The collected features are organized into a tree structure representing the relationship between the objects being reviewed and their components. The experimental results show that the proposed approach is able to automatically extract and identify the polarity for a large number of feature-sentiment expressions and achieve high accuracy.

The above models directly use terms as features to represent reviews and generate large sparse matrices owing to the huge number of features. To reduce the dimension of features, other scholars adopt the information of emotional polarities in texts as their features instead of simply representing texts by terms. For example, Yang, Song, and Tang (2013) studied Weibo messages and selected the frequencies of emotionally positive or negative words (as well as phrases and kaomojis), negatives, transitional words, gradable adverbs, special sentences and contextual keywords as features. Manek et al. (2017) proposed a statistical method using weight by Gini Index method for feature selection in sentiment analysis. More specifically, they extracted the terms as well as the opinion oriented words as features. There are also several approaches that combine two types of features. For example, Zheng, Wang, and Gao (2015) took Chinese online reviews as the research object, selected N-POS-grams (employ word n-grams plus POS n-grams as sentimental features) and N-char-grams as features to characterize the sentiment text. Akhtar et al. (2017) presented a cascaded framework of feature selection and classifier ensemble for sentiment analysis, and compact set of features performs better compared to the model that makes use of the complete features.

In general, feature selection and weighting are usually performed separately. Thus, the relationship between them is often ignored. Different feature selection approaches and weighting algorithms both affect the precision of the classification. Currently, one of the foci in sentiment analysis is to combine feature selection and the weighting method to obtain better classification performance. For example, Xu and Luo (2015) proposed an improved algorithm of feature weighting that takes into consideration the distribution information (DI) of inter-categories and intra-categories, as well as low frequency but high weight features. In addition, they combined the algorithm with TF-IDF to measure feature weighting. The proposed method obtains higher precision than TF-IDF does. Zhang et al. (2016) proposed two adaptive feature weighting approaches for naive Bayes text classifiers. One is the gain ratio-based feature weighting approach, they assume that all features only have two values of zero and nonzero and define the gain ratio of each feature partitioning a collection of training documents. The other is decision tree-based feature weighting for text classification, they set the weight of a feature to $1 + \lambda / \sqrt{d}$ (λ is a user-provided positive integer that determines how heavily to weight $1/ \sqrt{d}$) if the minimum depth at which the feature is tested in the built tree is d, and 1 if the feature does not appear in the built tree. The experimental results on a large number of text classification datasets validate their effectiveness and efficiency in terms of classification accuracy and elapsed training time in seconds, respectively. Parlak and Uysal (2018) compared two different feature selection methods namely Gini index and distinguishing feature selector and two different term weighting methods namely TF and TF-IDF in two pattern classifiers. Experimental results show that the most successful setting is the combination of Bayesian Network classifier, distinguishing feature selector, and TF term weighting method. These studies solve the problem of an uneven distribution among categories by combining feature selections and weighting.

2.2 Rule-based approaches

Statistics-based approaches usually depend on some corpora with manual annotations, and the performance is not good on a small corpus. To obtain better performance, we need to increase the number of training corpora to promote classification performance. In contrast, the rule-based approaches have little dependency on manually annotated corpora. With some domain knowledge, the classification rules can be artificially created. Compared with statistics-based ones, rule-based methods eliminate training of classification models and thus obtain results more quickly. For example, Endo, Saito, and Yamamoto (2006) leveraged dependency parsing to extract language identifiers that have a relatively large probability of co-occurrence with sentiment words. These researchers utilized “のが” (noga) (auxiliary word) and “ことが” (kotoga) (auxiliary word) as a possible identification of emotional expression to extract candidate statement blocks, filtered out the blocks by speech and ultimately obtained corresponding objects or reasons of emotional expressions. The precision of this method reached 45%. However, under the situation where the emotional expressions are missing, the corresponding objects and reasons cannot be correctly extracted. Siddiqua, Ahsan, and Chy (2017) combined a rule-based classifier with weakly supervised Naïve-Bayes classifier. Specifically, they introduce a set of rules for the rule-based classifier based on the occurrences of emoticons and sentiment-bearing words, whereas several sentiment lexicons are applied to train the Naive-Bayes classifier. Asghar et al. (2017) interated emoticons, modifiers and domain specific terms to analyze the reviews posted in online communities. Firstly, they use emoticon classifier to detect and score the emoticons. Secondly, they perform classification and score the modifiers and negations using a set of positive and negative modifiers and negation list. Thirdly, they apply sentiment classification of words using SentiWordNet-based classifier. Fourthly, they detect the domain specific words and label them with correct sentiment class and score. Finally, they perform sentiment classification of reviews at sentence and review level. The experiments obtain classification results with improved accuracy, precision, recall and F-measure as compared to comparing methods.

Regarding rule-based approaches, it is necessary to grasp certain knowledge and transfer it into classification rules. If some knowledge cannot be described as rules, the classification performance will usually decrease. In this case, the statistics-based approaches perform better than rule-based ones.

In recent years, much research has been conducted to combine the rule-based methods with statistics-based ones. For example, Xia et al. (2016) built a hybrid model that employs rules and statistical methods to detect explicit and implicit polarity shifts respectively. They propose a polarity shift elimination method to remove polarity shift in negations. Finally, they train base classifiers on training subsets divided by different types of polarity shifts, and use a weighted combination of the component classifiers for sentiment classification. The results on a range of experiments illustrate that the approach significantly outperforms several alternative methods for polarity shift detection and elimination. Also, inspired by the traditional methods like questionnaire, observation, and face-to-face interviews, Zhang and Zhou (2018) conducted an online investigation via automatic question answering. The researchers generate questions based on question templates, and extract corresponding answers using sentiment lexicon-based aspect-level sentiment analysis. Comparing the results with those obtained via traditional questionnaires, it shows that users' attitudes analyzed by AQA method are consistent with those extracted from traditional questionnaires. At the same time, much research has been conducted in multilingual and multi-domain settings. For example, a multilingual probabilistic topic modeling is proposed to train the classifiers on English–Italian, English–French and English–Spanish Wikipedia, respectively (Vulic et al., 2015). Stacked auto encoders are utilized to learn language-independent high-level feature representations on English–Chinese sentiment classification tasks of multiple data sets (Zhou et al., 2016). In their experiment, sentiment classifiers trained on the source language are leveraged to predict the sentiment polarity of the target language and the efficacy of the proposed cross-lingual approach is proved. Fernández, Esuli, and Sebastiani (2015) proposed a Distributional Correspondence Indexing method for domain adaptation in sentiment classification of English product reviews taken from Amazon.com with respect to four domains, i.e. books, DVDs, electronics and kitchen. The proposed method extracts term representations in a vector space common to source and target domains and obtains good performance.

Most prior research on sentiment analysis has been carried out on English and other European languages. To date, there have been few studies concentrating on Japanese reviews. In this study, we conduct sentiment analysis of Japanese reviews by utilizing statistics- and rule-based methods. Three groups of novel methods are proposed combining feature selection functions and the traditional TF-IDF method. Additional sentiment words and emoticons are chosen to complement the original emotional dictionary and an extensive comparative study is performed.

3 Datasets and Methods

3.1 Datasets

We collected reviews of Chinese tourist attractions from a Japanese travel website named 4Travel[2]. We selected 26 popular tourism cities in China, including Xiangfan, Wuhan, Wudang Mountain, Jingzhou, Beijing, Shanghai, Xi'an, Shaoxing, Ningbo, Hangzhou, Jiuzhaigou, Huanglong, Mount Emei, Chengdu, Xiamen, Suzhou, Nanjing, Kunshan, Mount Huangshan, Kaiping, Guangzhou, Lijiang, Shenyang, Dalian, Taipei, and Hong Kong. Based on all the reviews about the 26 cities, 3,069 reviews were randomly selected based on the number of the reviews about each city to form our dataset. Duplicate sentences were removed. The remaining sentences have different lengths varying from 4 to 79 words. We annotated the corpus manually and divided polarities into positive, neutral and negative. The numbers of sentences under each category are 1,545, 1,181, and 343, respectively. We divided the corpus into training and testing sets. The numbers of sentences under each category in the training set are 1,234, 959, and 262, respectively. The numbers of sentences under each category in the testing set are 311, 222, and 81, respectively. In the training and testing sets, the frequency distributions among different categories are uneven.

In the study, we utilized a Japanese word segmentation software named Mecab[3] for word segmentation and retain lemmas. After obtaining the results of word segmentation, we removed the meaningless numbers and punctuation marks.

3.2 Methods of feature selection

Generally, using terms as features to represent texts leads to a high dimension in sentence representation. To improve the processing efficiency and classification

[2] http://4travel.jp/overseas/area/asia/china/
[3] http://code.google.com/p/mecab/

performance, features need to be filtered to reduce dimensions and remove noise. We selected three functions, i.e. Chi-square (CHI), Information Gain (IG) and Mutual Information (MI), to conduct a comparative study.

3.3 Improved TF-IDF

Traditional TF-IDF does not consider the distribution of features in categories. After obtaining the results of feature selection, the relationship between retained features and categories cannot be embodied by TF-IDF. Zhang et al. (2015) found that the combination of TF, IDF and feature selection functions can achieve better classification results. Inspired by this study, we improved the traditional TF-IDF by combining TF, IDF and feature selection functions with the following steps. First, calculate the feature selection function values by utilizing CHI, MI and IG, separately. Second, rank the features in descending order in terms of the feature selection function values. Third, choose the top k features to achieve the goal of feature selection. Finally, in terms of feature weighting, multiply feature selection function values and TF-IDF and normalize the product. In this way, the document-feature matrix is obtained.

3.4 The sentiment dictionary and rule-based sentiment classification

3.4.1 Sentiment classification

Akira Nakamura's *Kanjo Hyogen Jiten* (emotional expression dictionary) (Nakamura, 1979) divides the emotional vocabulary into ten categories, i.e. "good", "sad", "tiresome", "shocked", "scared", "excited", "angry", "ashamed", "happy" and "calm". Among them, "good", "happy" and "calm" are regarded as positive categories, and "sad", "tiresome", "scared", "angry" and "ashamed" are regarded as negative ones. Most words in the sentiment dictionary can be assigned to one of emotional categories. Emotional categories of words, such as "shocked" and "excited", depend on the context. Thus, we cannot simply assign these words to the positive category or negative category.

3.4.2 Construction of the sentiment dictionary

Sentiment analysis depends on emotional dictionaries to a large extent. In this paper, construction of the emotional dictionary consists of the following parts.

1) Akira Nakamura's *Kanjo Hyogen Jiten* (KHJ) (Nakamura, 1979).
2) The Emotional Vocabulary in the field of Tourism (EVT). According to the characteristics of the tourism, terms with evident emotional polarities are extracted from the collected reviews and incorporated into the sentiment dictionary.

3) Kaomojis (emoticons). Kaomojis, a type of special symbol and an important basis for sentiment classification, can intuitively reflect reviewers' emotional tendencies. In addition, according to the positive and negative sentiment of expression, kaomojis can be divided into positive and negative.
4) Related Emotional Words (REWs). The emotional polarity of a REW is related to the polarity of the context. When REWs modify different things, their emotional polarity may change. In KHJ (Nakamura, 1979), words in the emotional categories of "shocked" and "excited" belong to REWs. For example, with "驚く (odoroku) (surprised) " in the sentence "この景色の美しさに驚いた (kono keshiki no utsukushisa ni odoroita) (The scenery is surprisingly beautiful.)", "驚く" expresses the surprise of "景色の美しさ (keshiki no utsukushisa) (the beautiful scenery)". Thus, it is a positive word here. In contrast, "驚く" expresses surprise at local congestion and clutter in the sentence "この混雑に驚いた (kono konzatsu ni odoroita) (It is surprisingly crowded.)", Thus, it is a negative word. In this study, words under the emotional categories of "shocked" (132 words in total) and "excited" (291 words) in KHJ (Nakamura, 1979) are considered as REWs. We also identify 62 REWs in the investigated dataset.
5) Terms with Implicit Emotional Tendency (TIETs). In the Japanese language, some terms themselves have certain emotional tendencies. We call them terms with implicit emotional tendency (TIETs). When a TIET expresses a certain type of sentiment, in most cases, it does not co-occur with emotional words of opposite polarity. For example, "のに (noni) (though, although, even though, in spite of, when, while or for)" usually expresses emotions such as dissatisfaction, blame, regret, repentance and the like. In the sentence of "色んな苦労したのに、天気のせいでいい景色をたのしめなかった (ironna kuroshita noni, tenki no seide ii keshiki wo tanoshimenakkata.) (Spending a hard day, we could not enjoy the beautiful scenery because of bad weather.)", "のに" is a clue to determine the emotional polarity. However, these types of terms have more than one usage. Thus, their usage needs to be examined. For example, "のに" is used to express the purpose of behavior in a situation such as "パスポートを取るのに、申請書を提出するのは必要だ (pasupooto wo toru noni, shinseisyo wo teisyusuru nowa hitsuyou da.) (Getting passport needs submit applications.)". After examining the dataset, we identify 13 TIETs, among which two terms are positive ones and eleven terms are negative ones.

3.4.3 Determination of classification rules

Classifying different types of emotional words in the sentiment dictionary requires various classification rules. The sentiment dictionary of KHJ (Nakamura, 1979), kaomojis and EVT are called Basic Emotional Vocabularies (BEVs). Transitional words and negative words are leveraged to judge the emotional polarity of sentences that contain these words more accurately. As for REWs, TIETs, transitional words and negative words, there are different processing rules.

1) Determination rules of BEVs
   Determine the frequencies of positive words and negative words in a sentence. If a sentence has more positive words than negative words, the sentence is assigned to the positive sentiment category and vice versa. If the numbers of positives and negatives are the same, the polarity of the sentence is considered to be neutral.
2) Determination rules of REWs
   Conduct dependency parsing on sentences that contain REWs and extract the chunks that depend on chunks embodying REWs. Next, determine the polarities of REWs based on whether these chunks contain BEVs. In particular, if there is a BEV in the chunk, assign the polarity of the REWs with the polarity of the BEV in the nearest chunks; otherwise, the polarity of the REWs is set to be neutral.
3) Determination rules of TIETs
   TIETs stand for emotional tendencies, and their emotional polarities hinge on whether the chunk before them contains BEVs with opposite emotional tendencies. Specifically, extract the chunk before TIETs and judge whether the terms with opposite emotional tendencies appear. If they do appear, the polarities of TIETs are neutral. Otherwise, TIETs maintain their original polarities.
4) Determination rules of transitional words
   Extract the transitional words and retain the clauses behind them. The emotional polarities of the original sentences are determined by the retained clauses.
5) Determination rules of negative words
   Upon the results of dependency parsing, extract the chunk containing negative words and the chunk ahead of them. Identify the object of negative words by considering the two chunks. If there are positive words in the two chunks, turn the polarity of the chunks into negative. If there are negative words, turn the polarity of the chunks into neutral.
6) Determination rules of sentences' polarities
   After setting the determination rules, the order of these terms needs to be arranged to determine the emotional polarities of the whole review.

The overall procedure is described as shown in Figure 1.

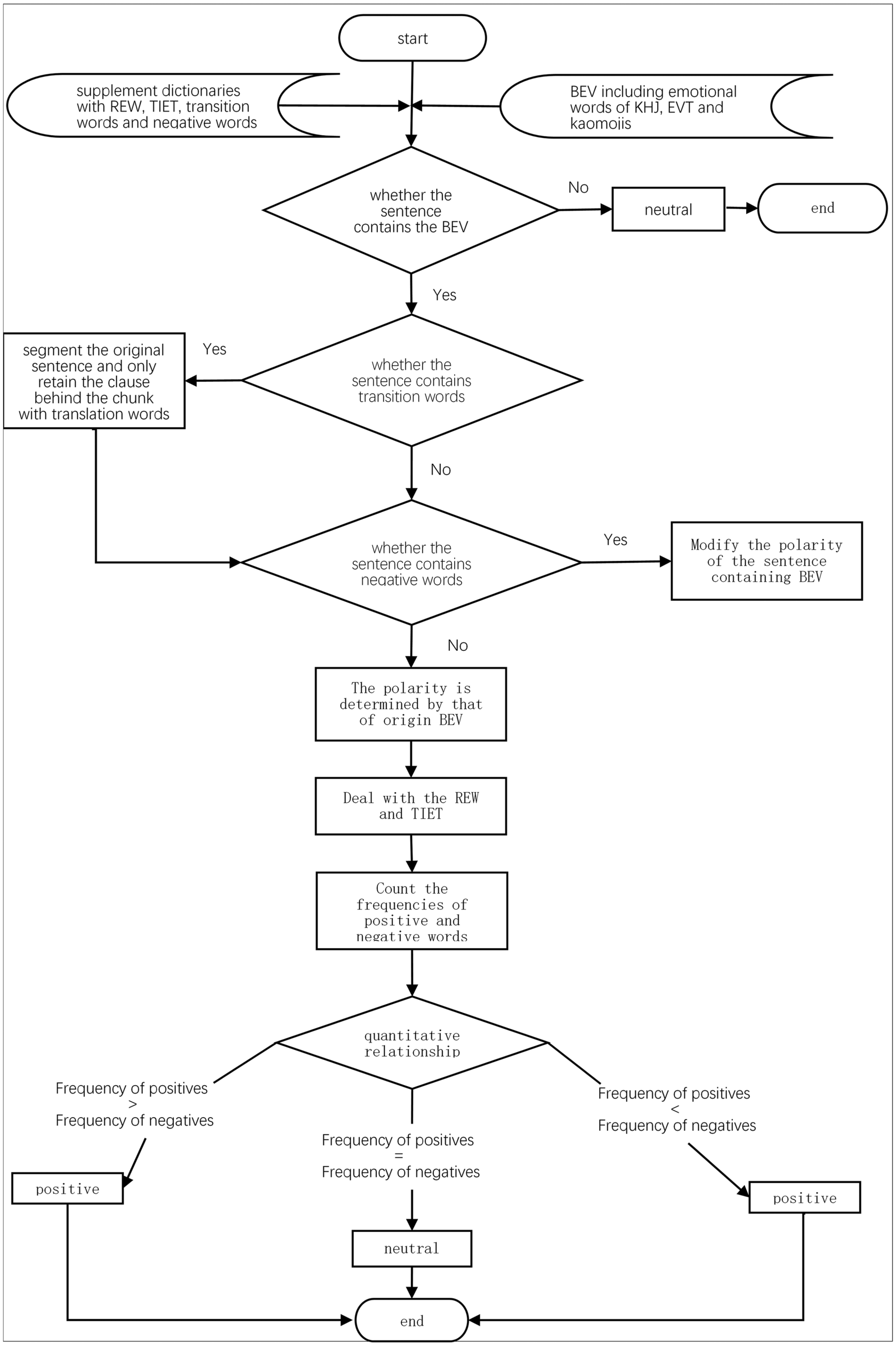
start
supplement dictionaries with REW, TIET, transition words and negative words
BEV including emotional words of KHJ, EVT and kaomojis
whether the sentence contains the BEV
No
neutral
end
Yes
whether the sentence contains transition words
Yes
segment the original sentence and only retain the clause behind the chunk with translation words
No
whether the sentence contains negative words
Yes
Modify the polarity of the sentence containing BEV
No
The polarity is determined by that of origin BEV
Deal with the REW and TIET
Count the frequencies of positive and negative words
quantitative relationship
Frequency of positives > Frequency of negatives
Frequency of positives = Frequency of negatives
Frequency of positives < Frequency of negatives
positive
neutral
positive
end

Figure 1. The flow diagram of the emotional polarity determination process.

The process of emotional polarity determination is as follows.

(1) If the sentence does not involve BEVs, the polarity is neutral and the polarity determination process ends. Otherwise, find whether the sentence contains transitional words. If yes, segment the original sentence and only retain the clause behind the chunk with transitional words (see the determination rules of transitional words).

(2) If the sentence consists of BEVs and does not include negative words, its polarity is determined by that of the original BEVs. If the sentence contains both BEVs and negative words, modify the polarity of the sentence according to the determination rules of negative words.

(3) Address REWs and TIETs according to the corresponding rules (see the determination rules of REWs and TIETs).

(4) Count the frequencies of positive and negative words in a sentence and assign the sentence to the sentiment category with higher frequencies of terms corresponding to the sentiment category. If the frequencies of positives and negatives are the same, the polarity of the sentence is neutral.

3.5 Evaluation metrics

Considering the uneven distribution of categories in the datasets, we use macro F1 (Macro-average) and micro F1 (Micro-average) as the evaluation metrics. F1 is calculated as shown in Equations (1), (2), and (3).

$$F1 = \frac{2 \times R \times P}{R + P} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

In Equation (1), R and P represent recall and precision, respectively. TP represents the number of sentences that are true positive, FP represents the number of sentences that are false positive, and FN represents the number of sentences that are false negative. The macro-average is obtained by first calculating the F1 value of each category then computing the mean value of F1. It is mainly influenced by small categories (e.g. the emotionally negative category). Micro-average accumulates the TP, FP and FN values of each category and calculates the F1 value. It is largely affected by large categories (e.g. the emotionally positive category). Due to the imbalanced data, macro-average more accurately reflects the performance of classifiers on negative categories, whereas the micro-average presents the performance on positive categories.

4 Results and Discussion

The experiment consists of two parts. One is based on methods of statistics and machine learning. The other is based on methods of dictionaries and rules.

4.1 The statistics-based methods

We utilized the original TF-IDF as the baseline method. The Support Vector Machine (SVM) was chosen as our classifier and ten-fold cross validations were conducted. The penalty parameter C is set to 100 and the kernel function of SVM is the Laplace kernel. The macro-average and micro-average values for the baseline method are 42.12% and 56.03%, respectively.

To compare with the baseline method, we designed the following study. 1) CHI, IG and MI are utilized as feature selection functions, and the numbers of features are set to 500, 1,000, 2,000, 3,000, 4,000, and 5,396, respectively. 2) Feature weightings are combined with the feature selection functions and the values of CHI*TF-IDF, IG*TF-IDF and MI*TF-IDF are normalized as the weights of features. 3) Tenfold cross-validation is conducted, and the values of penalty parameter C are set to 100, 200, and 300, respectively. The kernel functions of SVM are set to Gaussian, linear, polynomial, Laplace and tanh kernels, respectively. 4) Macro-average and micro-average are utilized to measure the performance.

The results of the comparative study are shown in Figure 2. The left column shows the results of the macro-average, and the right column demonstrates those of the micro-average. In each sub-figure, the horizontal axis represents the number of features and the vertical axis represents the F1 values.

Comparing the results of the baseline method with the results shown in Figure 2, it can be observed that all three combined methods (CHI*TF-IDF, IG*TF-IDF and MI*TF-IDF) outperform the baseline method in terms of both the macro-average and micro-average values. Thus, the classifier's performance can be highly improved by combining the feature selection functions and weightings. Simple but effective, the proposed combined method can resolve the problem of traditional TF-IDF lacking connections between emotional categories and features.

Comparing the three rows in Figure 2, we can see that CHI obviously achieve higher macro-average and micro-average values than IG when CHI and IG are utilized as feature selection functions. This finding is consistent with the result in the research of Sharma and Dey (2012). Also, the IG achieves lower macro-average and micro-average values than MI. The normalized product of the function value and TF-IDF is taken as the input of SVM. It is observed that the F1 value of CHI*TF-IDF is lower than that of MI*TF-IDF, accompanied by a minor fluctuation. The increase in the number of features leads to noise because the CHI value of new

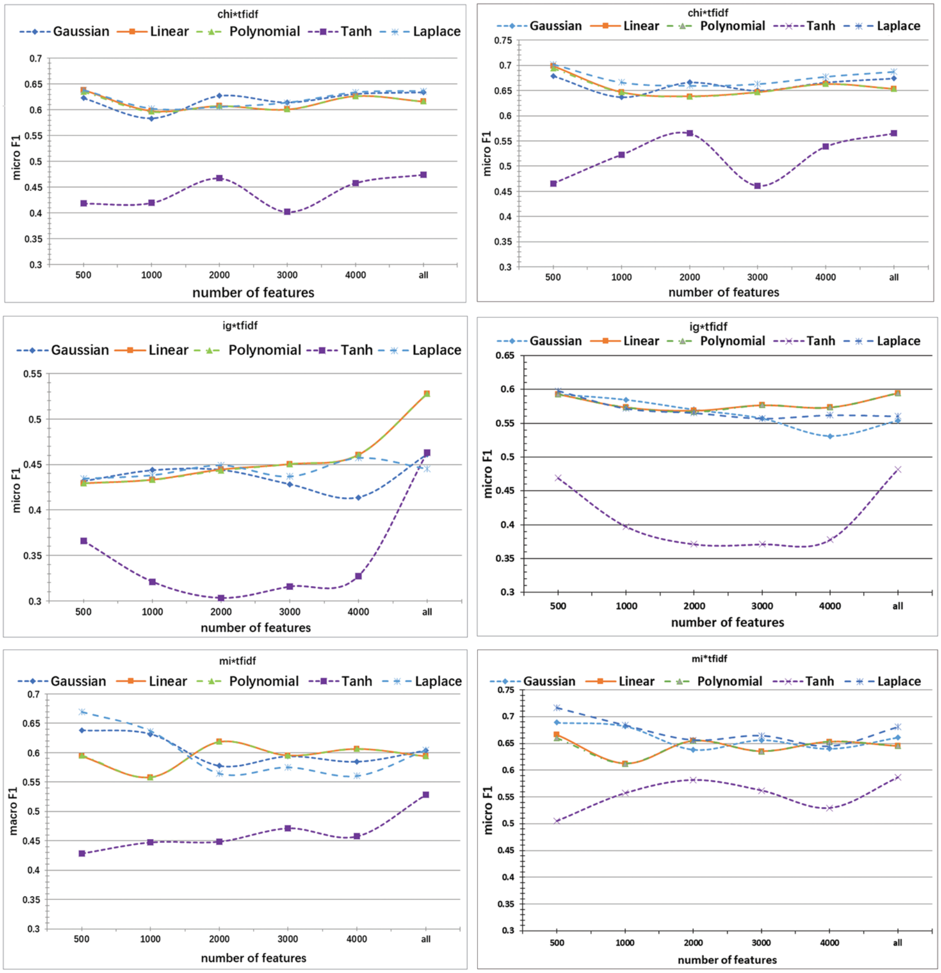

Figure 2. The results of different feature selection functions with different features and kernel functions.

features gradually decreases. It can be inferred from the small fluctuation range that CHI owns a stronger noise-tolerant ability.

With respect to the kernel functions, it is illustrated in Figure 2 that the performance varies among different numbers of features in the two evaluation metrics, i.e. the macro- and micro-averages. The Laplace kernel is outstanding because it achieves the highest values with a relatively stable fluctuation among the different numbers of features. The tanh kernel achieves the lowest macro-average and micro-average values and fluctuates the most. The performance of linear and polynomial kernels is basically the same. The Laplace kernel obtains the highest priority of recommendation in this study. Although it remains unclear whether the best kernel depends on the particular corpus, it can still be concluded that kernel functions have a large influence on the performance of sentiment classification of Japanese reviews.

We also compared the macro-average and micro-average values. It can be observed that the differences between the macro-average and micro-average values of the baseline and the methods in Figure 2 are more than 10 percents. In most cases, the macro-average values are significantly lower than the micro-average ones. The gaps are attributed to the uneven distribution of emotional categories and the intrinsic mechanism of the evaluation metrics. In terms of the macro-average, each emotional category is treated equally and thus, small categories affect the indicator to a great extent. As for the micro-average, each review is treated equally. Thus, large categories have a relatively large influence on the indicator.

To test whether other statistics-based methods could achieve better results, we also chose Decision Tree, K-Nearest Neighbor (KNN) and naïve Bayes as classifiers. Their macro-average values are shown in Figure 3.

It is found in Figure 3 that in the DT method, CHI*TFIDF achieves the best classification results followed by MI*TFIDF. The classification performance of the DT is greatly affected by the imbalance of categories. No data has been assigned to the small category of negative emotions for several times. In the KNN method, CHI*TFIDF also achieves the best classification results and IG*TFIDF does the worst. Different k values have been tried to achieve the highest macro-average values for each feature selection function. In Figure 3, the k values of the KNN are 7, 3, and 3 for CHI*TFIDF, IG*TFIDF and MI*TFIDF, respectively.

The Naïve Bayes classifier fails to achieve sound classification results on the data sets. In several experiments, it fails to classify the data, i.e. all the data are assigned to one category. As for the dimension of features, the Naïve Bayes is not good at classifying samples with high dimensional features. With the increasing number of features, the classifier’s performance decreases.

Comparing Figure 2 and Figure 3 reveals that the SVM classifier achieves better classification results than the Decision Tree, KNN and Naïve Bayes. The results confirm the findings of Sharma and Dey (2012) and Oma et al. (2014). In the former study, researchers explore the applicability of five feature selection methods (Document Frequency, IG, Gain Ratio, CHI, and Relief-F) and seven machine learning based classification techniques (Naive Bayes, SVM, Maximum Entropy, Decision Tree, KNN, Winnow, and Adaboost) for sentiment analysis on online movie reviews dataset. The experimental results show that CHI achieves better performance than IG for feature selection and SVM performs better than other techniques for sentiment-based classification. In the latter study, researchers present an empirical comparison of three classifiers (SVM, Naïve Bayes, and KNN) for Arabic sentiment classification. The experimental results indicate that the SVM

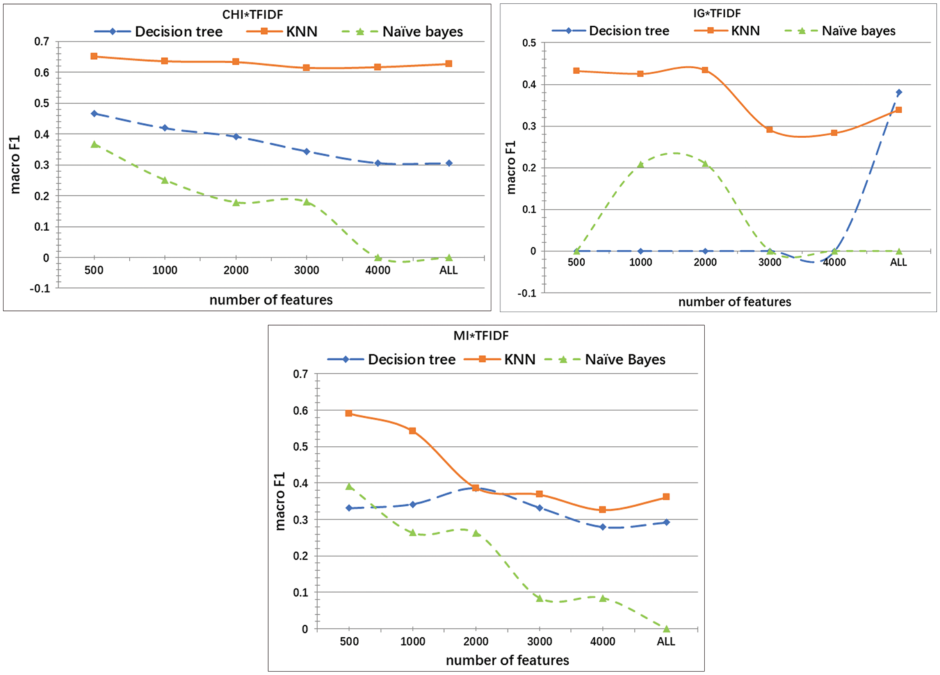

Figure 3. The results of different feature selection functions with different features and classifiers.

classifier outperforms the other techniques for Arabic sentiment-based classification. However, Palaniappan. T and Sundaraj (2014) obtained different results from us. They compared the performance of SVM and KNN classifiers in diagnosis respiratory pathologies using respiratory sounds from R.A.L.E database. The KNN classifier achieved better classification accuracy than SVM. This may be due to the differences between the application areas and the specific classification tasks.

4.2 Rule-based methods

To measure the influence of different rules on sentiment classification, the following seven groups of experiments were designed to conduct the comparative study. The first group of experiments is chosen as the baseline. 1) In the baseline method, KHJ (Nakamura, 1979) is utilized to directly classify sentences. 2) EVTs are complemented to the baseline method. 3) Kaomojis are added to the 2nd method. 4) REWs are supplemented to the 3rd method. 5) TIETs are appended to the 4th method. 6) Negative words are inserted to the 5th method. 7) Transitional words are introduced in the 6th method.

The macro-average and micro-average values are shown in Table 1. The precision, recall and F1 values of each category are demonstrated in Table 2.

Table 1. The results of rule-based methods.

	macro-average	micro-average
Group 1	0.4606	0.5350
Group 2	0.6419	0.6891
Group 3	0.6436	0.6924
Group 4	0.6436	0.6924
Group 5	0.6422	0.6914
Group 6	0.6527	0.6940
Group 7	0.6847	0.7185

Table 2. The classification results of all emotional categories.

	Positive			Neutral			Negative		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Group1	0.8366	0.3282	0.4714	0.4596	0.8967	0.6077	0.4780	0.2216	0.3028
Group2	0.7691	0.7547	0.7618	0.6429	0.6494	0.6462	0.5056	0.5306	0.5178
Group3	0.7713	0.7618	0.7665	0.6478	0.6494	0.6486	0.5042	0.5277	0.5157
Group4	0.7713	0.7618	0.7665	0.6478	0.6494	0.6486	0.5042	0.5277	0.5157
Group5	0.7713	0.7618	0.7665	0.6475	0.6469	0.6472	0.4986	0.5277	0.5127
Group6	0.7790	0.7508	0.7647	0.6426	0.6562	0.6493	0.5214	0.5685	0.5439
Group7	0.8057	0.7625	0.7835	0.6539	0.6926	0.6727	0.5871	0.6093	0.5980

Compared with the results of the baseline method, it is found that group 2 obviously exceeds the baseline in terms of F1 values in all three categories, i.e. positive, negative and neutral. Thus, it is more efficient to utilize EVTs to predict the emotional categories of sentences than to use the baseline. To explore the causes of the large differences (18.13% in macro-average and 15.41% in micro-average) between the two methods, we further investigate the items of KHJ (Nakamura, 1979) and find that KHJ (Nakamura, 1979) is mainly confined to the written language and lacks of colloquialisms, which leads to poor performance on online information reviews because the latter contains more colloquialisms and fewer formal written expressions than the former. The problem is well resolved in the EVTs as a large number of oral emotional words are supplemented.

According to Table 1, group 3 has achieved a less than 1% improvement compared with the results of group 2. The improvement is tiny because the size of the kaomoji vocabulary is relatively small (84 symbols) and just a small percentage of sentences (less than 3%) contains these kaomojis.

As seen in Table 1, the macro-average and micro-average values of groups 3 and 4 are exactly the same. We attribute the high approximation to the following. 1) In group 4, BEVs are utilized to identify the polarity of REWs, which leads to a large number of overlapping results. If the sentence does not contain BEVs, the REWs are considered to be neutral. We find that in most cases, neutral is not the correct category. 2) If the sentence contains only BEVs of a single emotional polarity or the frequencies of two polarities of BEVs are largely different, the polarity of the sentence is determined without any controversy. Thus, the existence of REWs has little influence on the ultimate decision on polarities. 3) The sentences that contain the REWs only accounts for 1.1% of all the reviews and approximately 0.22% of sentences contain both REWs and BEVs. In other words, only 0.88% of sentences that contain REWs are considered to be neutral. These sentences acquire the same category, i.e. neutral, in groups 3 and 4.

Comparing the results in group 4 and 5, it is found that the F1 values decrease by less than 1%. The reasons for the decrease are summarized as follows. 1) The emotional polarities of TIETs are determined by whether BEVs of opposite polarities exist within the context of two chunks. The out of vocabulary (OOV) problem of BEVs results in a small percentage of misjudgments. 2) Through an investigation into the corpus, we find that the number of reviews containing negative TIETs is much larger than that of positive. The positive-to-negative ratio is close to 1:69. The increase of negative words is larger than that of positive ones. When the negative TIET does not co-occur with the positive BEV, it indirectly results in an increase in the negative BEV. Thus, the precision of the negative category decreased from 50.41% to 49.86%. In addition, the precision of the neutral category decreased from 64.78% to 64.75%.

Comparing the results of groups 5 and 6, it can be found that the macro- and micro-average values improve 1.04% and 0.26%, respectively. It is inferred that negative words can affect the polarity of BEVs. On account of the determination rules of negative words (see Section 3.4.3), the number of sentences that are classified as negative or neutral increases. Thus, the recall value of the negative (or neutral) category is improved from 52.76% to 56.85% (or 64.69% to 65.62%).

Comparing the results of groups 6 and 7, it is found that the precision, recall and F1 values in all three categories are obviously improved (the increase ranging from 1.13% to 6.57%) when the transitional words are complemented in the vocabulary. It can be inferred from the large improvement that transitional words play an important role in determining the polarities of sentences. This is consistent with the language habits; the emotional polarities behind transitional words are decisive to the whole sentence.

Based on the results of the above seven groups of experiments, it is only under the condition of complementing TIETs that the performance is worse. Under the condition of complementing REWs, the performance remains the same. To further examine the impact of TIETs and REWs, we conducted three additional groups of experiments, i.e. removing REWs in DisI, TIETs in DisII and both of them in DisIII. The experiment results are shown in Table 3.

It is illustrated in the first column (DisI) of Table 3 that the precision, recall and F1 values in all three categories, i.e. positive, neutral and negative, are the same as the results of group 7. This finding further strengthens the evidence that REWs have no effect on the performance. It is demonstrated in the second column (DisII) that the precision and recall values in the neutral category and the precision in the negative category attain a small increase. It is inferred that TIETs have negative effects on the performance, specifically in the negative and neutral categories.

4.3 Comparison between the statistics-based and rule-based methods

To compare the overall performance of the statistics-based and rule-based methods, we utilized the optimal macro-average and micro-average values of the two methods. It can be observed from Tables 1 and 2 that the difference between the macro-average values of the two methods (0.6691602 and 0.6858637) is relatively small, as is the difference between the two micro-average values (0.7166124 and 0.7191268). The rule-based methods outperform the statistics-based ones with a narrow advantage. Comparisons between the statistics- and rule-methods contribute to moderating the controversy between the two types of methods, and strengthen the evidence that the rule-based ones may achieve higher performance if the determination rules are well designed. Theoretically, there is no necessity for manual annotations on corpora or additional training of the models for rule-based algorithms. The rule-based algorithms also have lower time complexities than statistics-based ones.

Table 3. The results of discarding words.

		DisI	DisII	DisIII
positive	precision	0.8057	0.8057	0.8057
	recall	0.7625	0.7625	0.7625
	F1	0.7835	0.7835	0.7835
neutral	precision	0.6539	0.6539	0.6539
	recall	0.6926	0.6943	0.6943
	F1	0.6727	0.6735	0.6735
negative	precision	0.5871	0.5921	0.5921
	recall	0.6093	0.6093	0.6093
	F1	0.5980	0.6006	0.6006
macro-average		0.6847	0.6857	0.6859
micro-average		0.7185	0.7191	0.7191

4.4 Comparison between the proposed methods and methods based on word embeddings

In the experiments mentioned above, on the one hand, we compared the influence of different feature selection functions, feature weightings and kernel functions on the performance of the emotional classification methods. On the other hand, we explored the influence of different rules on the performance of the emotional classification methods by applying 7 groups of rules mentioned in Section 4.2. To further verify the performance of the methods we proposed in this paper, we selected the best-performing statistics-based and rule-based method. For the statistics-based method, we chose SVM as classifier with Laplace kernel as its kernel function and CHI combined with feature weight as the feature selection function. For the rule-based method, we chose rule 7 (Group 7).We compared the two methods with existing sentiment analysis methods based on word embeddings, including the SVM (WE+SVM), Multi-Layer Perception (WE+MLP), Logistic Regression (WE+LR), Naïve Bayes (WE+NB), Decision Tree (WE+DT), Random Forest (WE+RF), Long-Short Term Memory (WE+LSTM) and Bi-directional Long-Short Term Memory (WE+BiLSTM). The Macro-average and Micro-average values of all these methods are shown in Table 4. It is worth noting that, the SVM, DT and NB methods in this section are different from the feature-based SVM, DT and NB in section 4.1. The three methods and the other methods in this section are based on word embeddings[④] pretrained on Wikipedia using Continuous Bag-Of –Words(CBOW) model with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives(Grave et al., 2018). To distinguish them, we add 'WE' as the prefix of all the word-embeddings-based methods.

According to Table 4, the statistics-based method we explored achieves the macro-average and micro-average values of 63.79% and 70.20% respectively. The rule-based method achieves 68.47% and 71.85%, all of which are better than the word embeddings-based methods. Regarding SVM, DT and NB, the three methods based on selected features in section 4.1 achieve better classification results than those based on word embeddings.

In addition, the results of the deep learning methods, i.e. LSTM and BiLSTM, are superior to other word-embeddings-based methods. However, there is still a large gap between these methods and the statistics-based and rule-based methods proposed in this paper. The reason may be attributed to the small size of the Japanese review corpus used in the paper. On a small dataset, the neural network may repeatedly try to catch information from the training data during the training process, resulting in inaccurate testing results. Another reason is that it is difficult to pretrain word embeddings on a small-size corpus. To address the problem, we use the word embeddings pretrained from the Japanese Wikipedia. The subsequent domain adaptation problem, i.e. from general domain to tourism domain, reduces the performances of LSTM and BiLSTM. Based on the analysis mentioned above, the optimal statistical method based on feature selection and rule-based method explored in this paper have achieved best sentiment classification performance.

[④] https://fasttext.cc/docs/en/crawl-vectors.html

Table 4. The classification results of different methods.

	Macro-average	Micro-average
WE+SVM	0.2241	0.5065
WE+MLP	02924	0.5277
WE+LR	0.3573	0.5619
WE+NB	0.3595	0.4121
WE+DT	0.3996	0.4723
WE+RF	0.4094	0.5326
WE+LSTM	0.4746	0.5668
WE+BiLSTM	0.4831	0.5602
CHI*TF-IDF (Laplace)	0.6379	0.7020
Rule-based (Group7)	0.6847	0.7185

5 Conclusions

In this paper, we conducted an extensive comparative study of statistics-based and rule-based methods of sentiment analysis of Japanese online tourism reviews.

In the statistics-based methods, we compared the influence of different feature selection functions, feature weightings, and kernel functions on the performance of the emotional classification methods. It is concluded that for the Japanese reviews, MI owns the highest ability of representation even with a small number of features, whereas CHI holds the strongest noise-tolerant ability. In general, the combination of feature selection functions and weightings can highly improve the overall performance. The kernel functions are demonstrated to be another key factor, and the Laplace kernel earns the highest reputation in this study. We experiment with different classifiers and find that the SVM classifier achieves better classification results than the Decision Tree, KNN and Naïve Bayes.

As for the rule-based methods, we conducted seven groups of experiments. It is concluded that EVT, kaomojis, negative and transitional words can obviously improve the performance in all three categories. The experiments also prove that REWs have no effect on the performances in all three categories, whereas TIETs have negative effects on the negative and neutral categories. Though the corpus is built upon tourism reviews, the discoveries still shed light on the intrinsic characteristics of the Japanese language and the influences on sentiment analysis due to the randomness of the data collection process.

The comparison between the rule-based methods and statistics-based methods indicates that the former performs better, exhibiting a narrow advantage, in the sentiment analysis of Japanese reviews. The results strengthen the evidence that the rule-based methods may achieve higher performance if the determination rules are well designed. Because these methods avoid the necessity of manual annotations on corpora or the additional training of the statistics models, the rule-based approaches gain higher priority for practical applications in this study. The findings can provide practical usage guidelines within the field of sentiment analysis of Japanese online tourism reviews and facilitate improving the service of tourist attractions.

In the future, more empirical studies will be conducted to further improve the classification performance and verify the validity of the proposed methods on other languages. Further studies will be carried out to address domain adaptation problem for deep learning-based methods.

Acknowledgments

This study is supported by the National Natural Science Foundation of China under the grant #71373286 and # 71603189 and the Major Project of the Ministry of Education of China (Grant No. 17JZD034).

Author Contributions

C.M. Yu (yucm@zuel.edu.cn) conceived and designed the analysis, collected and analyzed data, performed the analysis and drafted the paper. X.Y. Zhu (1693692423@qq.com) implemented the experiment and revised the paper. B.L. Feng (362205285@qq.com) revised the paper. L. Cai (chenvily@qq.com) revised the paper. A. Lu (anlu2009@whu.edu.cn) is responsible for the paper and the communication with the editorial department.

References

Abd-Elhamid, L., Elzanfaly, D., & Eldin, A.S. (2016). Feature-based sentiment analysis in online Arabic reviews. In Proceedings of 11th International Conference on Computer Engineering & Systems (pp.260–265). IEEE. doi: 10.1109/ICCES.2016.7822011

Akhtar, M. S, Gupta, D., & Ekbal, A. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. Knowledge-Based Systems, 125, 116–135. doi: 10.1016/j.knosys.2017.03.020

Asghar, M.Z., Khan, A., Ahmad, S., Qasim, M., & Khan, I. A (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PloS One, 12(2), e0171649. doi: 10.1371/journal.pone.0171649

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016) Enriching word vectors with subword information. arXiv:1607.04606v2.

Contratres, F.G., Alves-Souza, S.N., Filgueiras, L.V.L., & DeSouza, L.S. (2018). Sentiment analysis of social network data for cold-start relief in recommender systems. In Proceedings of World Conference on Information Systems and Technologies (pp.122–132). Springer, Cham. doi: 10.1007/978-3-319-77712-2_12

Endo, D., Saito, M., & Yamamoto. (2006).The extraction of emotional representation by using dependency relation. In Proceedings of Natural Language Processing.

Fernández, A.M., Esuli, A., & Sebastiani, F. (2016). Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. Journal of Artificial Intelligence Research, 55(1), 131–163. doi: 10.1613/jair.4762

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Impana, P., & Kallimani, J.S. (2017). Cross-lingual sentiment analysis for Indian regional languages (pp.1–6). In Proceedings of International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques.

Ma, W., & Deng, Y. (2013). New feature weighting calculation method for short text. Journal of Computer Applications, 33(8), 2280–2292.

Manek, A.S., Shenoy, P.D., Mohan, M.C., & Venugopal, K.R. (2016). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web, 20(2), 135–154. doi: 10.1007/s11280-015-0381-x

Nakamura, A. (1979). Kanjo Hyogen Jiten. Toukyouto Rokkou Press.

Omar, N., Albared, M., Al-Moslmi, T, &. Al-Shabi, A. (2014) A comparative study of feature selection and machine learning algorithms for Arabic sentiment classification. Information Retrieval Technology, 8870, 429–443. doi: 10.1007/978-3-319-12844-3_37

Parlak, B., & Uysal, A.K. (2018). On Feature weighting and selection for medical document classification. Developments and Advances in Intelligent Systems and Applications (pp. 269–282). Springer, Cham.

Palakvangsa-Na-Ayudhya, S, Sriarunrungreung. V, Thongprasan, P., & Porcharoen, S. (2011) Nebular: A sentiment classification system for the tourism business. In Proceedings of 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp.293–298). IEEE. doi: 10.1109/JCSSE.2011.5930137

Palaniappan, R., Sundaraj, K., & Sundaraj, S. (2014). A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signal. BMC Bioinformatics, 15(1), 223. doi: 10.1186/1471-2105-15-223

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2016). Multi-lingual opinion mining on YouTube. Information Processing and Management, 52(1), 46–60. doi: 10.1016/j.ipm.2015.03.002

Sharma, A., & Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In Proceedings of the 2012 ACM research in applied computation symposium (pp.1–7). ACM. doi: 10.1145/2401603.2401605

Siddiqua, U.A., Ahsan, T., & Chy, A.N. (2017). Combining a rule-based classifier with weakly supervised learning for twitter sentiment analysis. In Proceedings of International Conference on Innovations in Science (pp.1–4), Engineering and Technology. doi: 10.1109/ICISET.2016.7856499

Song, W., Cai, Y., Wu, B., & Sun, T. (2012). A new active learning strategy in nearest neighbor classifier. In Proceedings of the International Conference on Machine Learning and Cybernetics (pp.729–734). Xi'an, China. IEEE. doi: 10.1109/ICMLC.2012.6359015

Soni A K. (2017). Multi-lingual sentiment analysis of Twitter data by using classification algorithms. In Proceedings of 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp.1–5). doi: 10.1109/ICECCT.2017.8117884

Vulic, I., Smet, W.D., Tang, J., & Moens, MF. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. Information Processing & Management, 51(1), 111–147. doi: 10.1016/j.ipm.2014.08.003

Xia, R., Xu, F., Yu, J., Qi, Y. & Cambria, E (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. Information Processing & Management, 52(1), 36–45. doi: 10.1016/j.ipm.2015.04.003

Xiao, X., Lu, J., Yu, L., & Gong, H. (2015). Research on feature selection algorithm based on the lowest term frequency of CHI. Journal of Southwest University (Natural Science Edition), 37(6), 137–142.

Xu, F.Y., & Luo, Z.S. (2015). An improved approach to term weighting in automated text classification. Computer Engineering and Application, 4(1), 181–184.

Yang, W., Song, J.J., & Tang, J.Q. (2013). A study on the classification approach for Chinese MicroBlog subjective and objective sentences. Journal of Chongqing University of Technology (Natural Science), 27(1), 51–56.

Yang, Y.M., & Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning (pp. 412–420). Nashville, TN, USA.

Zhang, C.Z., & Zhou, Q.Q. (2018). Online investigation of users' attitudes using automatic question answering. Online Information Review, 2018, 42(3), 419–435. doi: 10.1108/OIR-10-2016-0299

Zhang, L. (2015) Aspect: eight summary of "Internet + tourism" industry trend in 2016. Retrieved from http://mi.chinabyte.com/299/13641299.html.

Zhang, L., Jiang, L., Li, C., & Kong, G. (2016). Two feature weighting approaches for naive Bayes text classifiers. Knowledge-Based Systems, 100, 137–144. doi: 10.1016/j.knosys.2016.02.017

Zheng, L., Wang, H., & Gao, S. (2015). Sentimental feature selection for sentiment analysis of Chinese online. International Journal of Machine Learning and Cybernetics, 9(1), 75–84.

Zhou, G.Y., Zhu Z.Y., He, T.T., & Hu, X.T. (2016). Cross-lingual sentiment classification with stacked auto-encoders. Knowledge and Information Systems, 47(1), 27–44. doi: 10.1007/s10115-015-0849-0

Zin, H.M., Mustapha, N., Murad, M.A.A. & Sharef, N.M. (2018). Term weighting scheme effect in sentiment analysis of online movie reviews. Advanced Science Letters, 24(2), 933–937.