Original Article

A method for discriminating travel reviews from commercial speech

Takahiro Hayashi (Faculty of Engineering, Niigata University, hayashi@ie.niigata-u.ac.jp)
Yoshio Seino (Graduate School of Science and Technology, Niigata University)

Abstract
This paper presents a computational method for discriminating travel reviews from commercial speech. Today many travel reviews can be found on the Web. Mining useful information from these reviews is important not only for consumers but also for companies like travel agencies. Travel reviews written by ordinary people are essentially different from commercial speech like advertisements generated by companies and individuals for the intent of making a profit. We propose a computational method for discriminating travel reviews from commercial speech. Assuming that subjective words often occur in travel reviews rather than commercial speech texts, we define the subjectivity score of each word in a document. Evaluating the total subjectivity score of a document using all the words in the document, the proposed method identifies whether the document is classified as travel review or not. From experiments, we have confirmed the proposed method can accurately classify documents into travel reviews and commercial speech texts.

Keywords
travel reviews, commercial speech, opinion mining, text mining, natural language processing

1. Introduction

Gathering and analysing travel reviews such as reputations about tours is an important task in a wide range of tourism-related applications. For example, analysing travel reviews help travel agency identify the potential revenues of tour products, judge the success of new tour packages, and improve the quality of customer service.

Owing to the importance of reviews on the Web, many opinion mining techniques have been developed [Ravi, 2015; Khan et al., 2014; Selvam and Abiami, 2009; You, 2016; Jyoti and Rao, 2016; Vinodhini and Chandrasekaran, 2012]. One of the major topics in opinion mining is to identify whether the opinion is positive, negative or neutral.

Travel reviews on the Web are often contaminated with commercial speech like advertisements. Commercial speech is generated by companies and individuals for the intent of making a profit. Therefore, eliminating commercial speech from documents is an important data cleansing process for opinion mining.

A computational method for discriminating personal web pages from non-personal web pages has been proposed [Hayashi et al., 2008; 2009]. Assuming subjective words such as “bad” are more likely to occur in personal web pages than non-personal pages, this method defines the subjectivity score of a web page by counting the number of subjective words used in the page, and regards pages having high subjectivity scores as personal web pages.

In this paper, we extend this previous method to solve the discrimination problem of travel reviews. We have found that the previous method has two problems. The first problem is that the number of subjective words used in the previous method was limited. This is because they were selected by human labour. The second problem is that the selected subjective words depend on Japanese. Therefore, it is difficult to simply apply the previous method to discrimination of texts written in other languages.

In this paper, we propose a language-independent method for discriminating travel reviews from commercial speech. In the proposed method, subjective words are automatically generated by analysing text sets (corpora) written in various languages. By this corpus-based approach, the above two problems can be solved.

Today, the main way for the surveys of personal opinions about travel tours is checks of reviews. However, the task of collecting personal reviews from the Web by hand still needs hard labour. The major contribution of this study is to provide an effective way for collecting personal reviews which are discriminated from commercial speech texts. Another important contribution is that the proposed method can be used as a data cleansing method for various opinion mining techniques. More valuable and reliable mining results can be obtained by the proposed method, which benefits to the field of tourism research.

The rest of the paper is organized as follows. In section 2, we describe the related works. In section 3, we describe the details of the proposed method. In section 4, we explain experiments using actual travel reviews and advertisement texts on the Web. In the section, we show that the proposed method can accurately discriminate travel reviews form commercial speech texts independently from languages. Finally, we conclude the paper in section 5.

2. Related works

A computational method for discriminating personal web pages from non-personal web pages has been proposed [Hayashi et al., 2008; 2009]. The previous method focuses on the appearance ratios of four kinds of subjective words: negative meaning expressions, sentence-final particles, interjections, and specific symbols such as emoticons. By using these subjective words, the subjectivity score of a given text is defined. In addition, based on the scoring model, the method judges texts having high subjectivity scores as personal web pages.

As explained in the previous section, the four kinds of subjective words were selected by human labour and these words are strongly dependent on the Japanese Grammar. Therefore, it is difficult to apply the method to other languages.

The method proposed in this paper is language-independent. In addition, the dictionary of subjective words is automatically generated by analysing two kinds of text sets: a set of travel review texts and a set of commercial speech texts.

Discriminating travel reviews from commercial speech can be regarded as a binary classification problem. The purpose of binary classification problems is to classify documents into two types of pre-defined categories. In the field of NLP (natural language processing), there are many sophisticated methods for document classification [Lie et al., 2015; Mikolov et al., 2010; Mikolov et al., 2013]. Many of them utilize machine learning techniques such as deep learning. These methods can solve complex classification problems such as multi-class classification and multi-label classification. However, these methods require a huge volume of training data. This requirement is a big problem for many travel-related applications because enough training data cannot be obtained in many cases.

On the other hand, our method in this paper requires neither a complex computational model nor a huge dataset. In order to find the difference between travel reviews and commercial speech, the proposed method just focuses on the difference of appearance ratios of each word between the two types of documents. Even though the approach is simple and straightforward, the proposed method can discriminate personal reviews from commercial speech with satisfactory accuracy (The efficiencies of the proposed method are shown in section 4). The simplicity and accuracy is an advantage of the proposed method.

3. Proposed Method

3.1 Definitions of travel reviews and commercial speech

The purpose of the proposed method is to discriminate travel reviews from commercial speech. Before explaining the details of the proposed method, in this subsection we define the words “travel reviews” and “commercial speech”.

In this paper, the word of “travel reviews” means review texts contains personal opinions about travel. The word of “personal opinion” means a comment described by a personal writer. Hence, a comment described by a professional writer is not regarded as a personal opinion. Therefore, review texts written by professional writers are not included in the “travel reviews”.

The reason why this paper distinguishes personal opinions from professional ones is there is the difference between minds of writers. Concretely, professional writers tend to emphasise positive comments about subjects for their sponsors and seldom report negative comments about them. On the other hand, personal writers report any comments about subjects. Therefore, for people who want to collect real public impressions about subjects, personal opinions would be more useful than professional opinions.

The word “commercial speech” means texts which are not classified as “travel reviews”. Reviews written by professional writers are classified as “commercial speech”.

3.2 Outline

In the proposed method, we assume that used words in reviews and commercial texts have different trends. For example, the word “disappoint” may occur in reviews, but it seldom appears in commercial speech. In this case, the word “disappoint” is helpful for discrimination of the two kinds of texts.

The proposed method measures the appearance ratios of each word in two kinds of text corpora $\mathrm{C_p}$ and $\mathrm{C_n}$. Corpus $\mathrm{C_p}$ is a set of travel review texts and corpus $\mathrm{C_n}$ is a set of commercial speech texts such as advertisements of tours.

Currently it is difficult to build a corpus using publicly available dataset because there are few such dataset. However, with the increase of tourism-related services, today there are many online review sites about travel tours. In this study, we build corpora by collecting reviews and advertisements about travel tours from the site of VELTRA (http://www.veltra.com), which is a portal site of a travel agency for promoting tours.

By subtracting the appearance ratios of each word between the two corpora, the subjectivity score of the word is defined. In the proposed method, words having high subjectivity scores are regarded as subjective words.

Since subjective words are expected to occur in travel reviews more frequently than commercial speech, travel reviews can be separated from commercial speech by focusing on these subjective words.

3.3 The subjectivity scores of words

As described in 3.1, the proposed method focuses on the difference of appearance ratios of each word between the two text corpora $\mathrm{C_p}$ and $\mathrm{C_n}$. Using the two corpora, the subjectivity score of word w is defined as

$$s(w) = s_p(w) - s_n(w),$$

where $s_p(w)$ and $s_n(w)$ are respectively defined as

$$s_p(w) = \frac{r_p(w)}{r_p(w) + r_n(w)},$$

$$s_n(w) = \frac{r_n(w)}{r_p(w) + r_n(w)}.$$

Here, $r_p(w)$ and $r_n(w)$ are the appearance ratios of word w in the corpus $\mathrm{C_p}$ and $\mathrm{C_n}$, respectively. The definitions of $r_p(w)$ and $r_n(w)$ are as follows:

$$r_p(w) = \frac{\sum_{d \in \mathrm{Cp}} f(w, d)}{\sum_{d \in \mathrm{Cp}} F(d)},$$

$$r_n(w) = \frac{\sum_{d \in \mathrm{Cn}} f(w, d)}{\sum_{d \in \mathrm{Cn}} F(d)}.$$

where $F(d)$ is the total count of words in document d, and $f(w, d)$ is the number of word w in document d.

Since subjective words like negative-meaning words are expected to be used more frequently in in review texts rather than commercial speech texts, $s_p(w)$ is expected to be larger than $s_n(w)$. As a result, the subjective score $s(w)$ becomes positive ($s(w) > 0$). In contrast, the subjective scores of words often occurring in commercial speech tend to be negative ($s(w) < 0$).

3.4 Elimination of extremely common and uncommon words

Eliminating extremely-common words, which are known as *stop words*, from a document is an important pre-processing for text mining because these words give little meaningful information. In the proposed method, words occurring in more than 95 % documents, are eliminated as stop words.

In addition, extremely few frequent words, such as misspelled words and coined words not very commonly used, give little information in document classification. Eliminating extremely uncommon words are important for reducing the computational cost and saving the memory. In the proposed method, words occurring in less than 3 times in all the documents in the two corpora are eliminated.

3.5 Discrimination of Travel reviews from commercial speech

Based on the subjective scores of each word, the total subjectivity score of document d is defined as follows:

$$S(d) = \frac{1}{|W(d)|} \sum_{w \in W(d)} s(w),$$

where $W(d)$ is the set of words in document d. If the total subjectivity score of document d is positive, i.e., $S(d) > 0$, the document is classified as a travel review; if otherwise, i.e., $S(d) \leq 0$, the document is classified as commercial speech.

4. Experiments

4.1 Experimental settings

In order to confirm the effectiveness of the proposed method, we conducted experiments. In the experiments, using actual travel reviews and commercial speech texts on the Web, we confirmed how the proposed method effectively distinguishes travel reviews from commercial speech.

As the first experiment, we used 20,000 documents written in Japanese (10,000 travel reviews and 10,000 commercial speech texts). As the second experiment, we used 20,000 documents written in English (10,000 travel reviews and 10,000 commercial reviews).

These documents were collected from the web site of VELTRA (http://www.veltra.com), a travel agency for promoting tours. As travel reviews, we collected the reviews posted by people who actually participated in the tour. As commercial speech texts, we collected the texts created by the travel agency for explaining the tour contents.

In order to split a Japanese text into a word sequence, we used Japanese part-of-speech analyser MeCab (http://taku910.github.io/mecab/).

We measured the classification accuracy of the proposed method with the 10-fold cross validation manner [Ron, 1995]. That is, the documents used in the experiments are divided into 10 subsets. By using 9 subsets as a training set, the subjectivity score of each word is calculated. After that, the documents in the remaining one subset is used as a test set for classification. By changing the combination of training and test sets, the classification experiments are repeated 10 times. The average classification performance across the 10 rounds is defined as the classification accuracy of the proposed method.

For each round, the performance of classification results is evaluated by the *rand-index* and *the precision/recall measures*.

The rand-index RI is a measure that evaluates how accurately the system classifies data into appropriate classes. The rand-index RI is defined as follows:

$$RI = \frac{TP + TN}{TP + FP + TN + FN},$$

where TP (true positive) is the number of documents classified as travel reviews and that are truly travel reviews, FP (false positive) is the number of documents classified as travel reviews and that in reality are commercial speech, TN (true negative) is the number of documents classified as commercial speech texts and that are truly commercial speech texts, and FN (false negative) is the number of documents classified as commercial speech and that in reality are travel reviews.

The precision/recall measures are widely used in the information retrieval field. Precision is a measure of how many documents classified as travel reviews were actual travel reviews. Recall is a measure of how many actual travel reviews were classified correctly. Precision P and recall R are defined as follows:

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN}.$$

4.2 Results

Table 1 and 2 show the results of the cross-validation when using Japanese and English documents, respectively. From the table, we can confirm that high classification performances (rand-index RI, precision P and recall R) were obtained both in Japanese and English documents. These results indicate that the proposed method can effectively discriminate travel reviews from commercial speech regardless of their languages.

Figure 1 and 2 shows the distributions of total subjectivity scores of the Japanese and English documents, respectively. As shown in these figures, travel reviews and commercial speech texts are clearly separated by the subjectivity score. These results show the effectiveness of the proposed model of subjectivity score.

Table 1: The results of 10-fold cross validation using Japanese documents

Round	RI	P	R	TP	TN	FP	FN
1	0.986	0.996	0.976	996	975	4	25
2	0.979	0.998	0.961	998	960	2	40
3	0.987	0.996	0.977	996	977	4	23
4	0.981	0.995	0.967	995	966	5	34
5	0.986	0.994	0.977	994	977	6	23
6	0.987	0.992	0.982	992	982	8	18
7	0.98	0.988	0.972	988	972	12	28
8	0.986	0.996	0.976	996	975	4	25
9	0.986	0.991	0.98	991	980	9	20
10	0.981	0.993	0.969	993	968	7	32
Total	0.984	0.994	0.973	9939	9732	61	268

Table 2: The results of 10-fold cross validation using English documents

Round	RI	P	R	TP	TN	FP	FN
1	0.976	0.969	0.983	983	969	31	17
2	0.988	0.983	0.992	992	983	17	8
3	0.985	0.977	0.992	992	977	23	8
4	0.978	0.97	0.986	986	970	30	14
5	0.984	0.976	0.992	992	976	24	8
6	0.983	0.974	0.992	992	973	27	8
7	0.985	0.979	0.99	990	979	21	10
8	0.985	0.982	0.988	988	982	18	12
9	0.981	0.972	0.99	990	972	28	10
10	0.985	0.98	0.989	989	980	20	11
Total	0.983	0.976	0.989	9894	9761	239	106

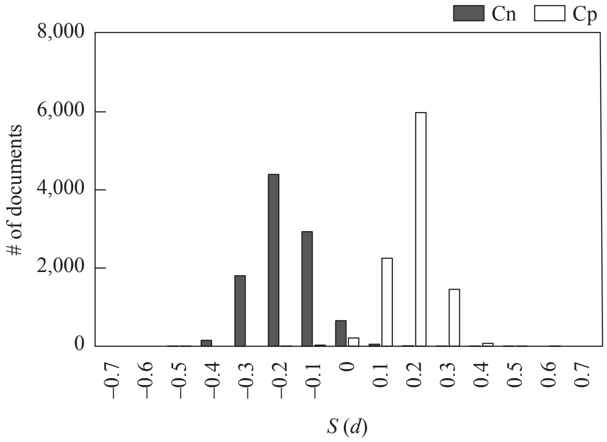

Figure 1: The distribution of total subjectivity scores of the Japanese documents in C_p (travel reviews) and C_n (commercial speech)

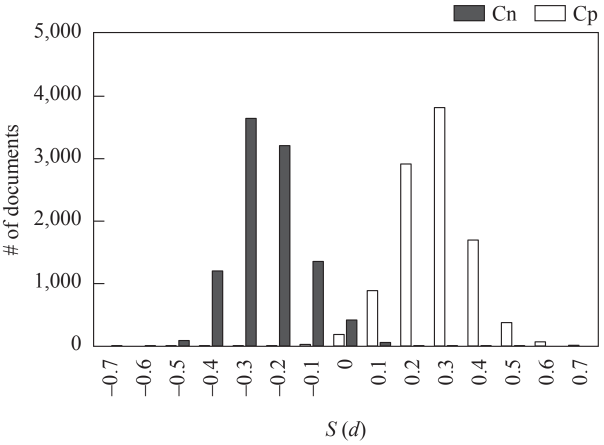

Figure 2: The distribution of total subjectivity scores of the Japanese documents in C_p (travel reviews) and C_n (commercial speech)

4.3 Analysis of frequent words in travel reviews and commercial speech

Table 3 shows frequent words in travel reviews and ones in commercial speech texts. Table 4 shows frequent words in English documents. In these tables, the subjectivity scores of the frequent words are also shown. As shown in these tables, subjective words have high subjectivity scores and that non-subjective words have low (negatively-high) subjectivity scores. These results indicate that by checking their subjectivity scores, subjective words can be clearly separated from other

Table 3: Japanese words having high subjectivity scores in travel reviews and commercial texts

Travel reviews	Commercial speech
笑 (lol)(0.99), ラッキー (lucky)(0.97),	最古 (earliest)(–0.99), ユネスコ (UNESCO)(–0.99),
ちょうど (just)(0.98), かなり (very)(0.97),	聖地 (holy site)(–0.98), 首都 (capital)(–0.98),
ビックリ (amazed)(0.96), きちんと (properly)(0.95),	国宝 (national treasure)(–0.98), 旅館 (hotel)(–0.98),
悪い (bad)(0.94), ひどい (terrible)(0.94),	古都 (ancient city)(–0.98), 味覚 (taste)(–0.98),
きつい (heavy)(0.93), 殆ど (almost)(0.94),	シェフ (chef)(–0.97), 避暑 (summer)(–0.97),
嫌 (dislike)(0.94), つらい (tough)(0.91),	プロデュース (produce)(–0.97), 産業 (industry)(–0.97),
ユーモア (humor)(0.90), 配慮 (0.89),	注目 (attention)(–0.97), 下町 (downtown)(–0.96),
改善 (improve)(0.88), メリット (merit)(0.85)	森林浴 (forest therapy)(–0.96),
	中世 (the medieval era)(–0.95),
	お客様 (customer)(–0.95), お子様 (kids)(–0.95),
	地名 (location name)

Table 4: English words having high subjectivity scores in travel reviews and commercial texts

Travel reviews	Commercial speech
I (0.99), my (0.99), bad (0.99), quite (0.99),	heritage (0.99), seasonal (0.99), site (0.98),
highly (0.99), guy (0.99), disappoint (0.98),	temple (0.98), discover (0.97), brewery (0.97),
absolutely (0.98), recommend (0.98), worth (0.98),	ritual (0.97), electric (0.97), regional (0.96),
extremely (0.97), nice (0.97), enough (0.96),	century (0.96), chapel (0.96), century (0.96),
cheap (0.96), awesome (0.95), outstanding (0.96),	countryside (0.95), seaside (0.94), era (0.94),
our (0.96), difficult (0.94), excellent (0.93),	symbol (0.93), city (0.90), eastern (0.90),
poor (0.93), loud (0.93), helpful (0.92)	classic (0.88), location name

words.

For example, as subjective words, negative-meaning expressions occur more frequently in travel reviews than commercial speech. On the other hand, neutral expressions related to tourism and location names often occurs in commercial speech texts.

In English documents, the first-person forms, e.g., “I” and “our”, often occur in travel reviews. These words seldom occur in commercial speech texts. The same trend would be found in other Indo-European languages such as French. Focusing on the usage of first-person forms in these languages would be helpful for discriminating personal texts from others.

We found sentence-final particles and specific symbols appearing at the end of a sentence such as emoticons have relatively high subjectivity scores. Sentence-final particles and emoticons express linguistic modality, resulting in the frequent use of them in personal texts. Since sentence-final particles are common concepts in Asian languages, these words would be useful for the discrimination problem of travel reviews written in any other Asian languages.

The importance of subjective words in the discrimination problem has already been shown in the previous study [Hayashi et al., 2008; 2009]. As explained in Sec. 2, in the previous study, subjective words are selected carefully by hand. However, since the number of subjective words is limited, it is difficult to obtain high classification performances. In contrast, in this study, the dictionary of subjective words can be generated automatically from corpora. Since a large number of subjective words can be found by the proposed method, high classification performances can be obtained as shown in the Table 1 and 2.

Further experiments for other languages including French, Chinese and Arabic remain as a future work.

5. Conclusion

In this paper, we have presented a language-independent method for discriminating travel reviews from commercial speech. The experimental results showed travel reviews are accurately discriminated from commercial speech. Further experiments using other languages including French, Chinese and Arabic remain as a future work.

The proposed method can be used as a data cleansing method for opinion mining. By combining the proposed method with other post processing techniques such as sentiment analysis, intelligent tour-related services can be developed.

References

Hayashi, T., Abe, K., and Onai, R. (2008). Retrieval of personal web documents by extracting subjective expressions. *Proceedings of 22nd IEEE Advanced Information Networking and Applications Workshops*, 1187-1192.

Hayashi, T., Abe, K. Abe, Roy, D., and Onai, R. (2009). Discrimination of personal web pages by extracting subjective expression. *International Journal of Business Intelligence and Data Mining*, Vol. 4, No. 1, 62-77.

Jyoti, R. S. (2016). A survey on sentiment analysis and opinion mining. *Proceedings of the International Conference on Advances in Information Communication Technology and Computing*, No. 53.

Khan, F. H., Bashir, S., and Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, Vol. 57, 245-257.

Liu, Y., Liu, Z., Chua, T., and Sun, M. (2015). Topical word embedding. *The 29th AAAI Conference on Artificial Intelligence*, 2418-2424.

Mikolov, T. Karafiát, M., Burget, L., Cernocký, J., and Khudanpur. S. (2010) Recurrent Neural Network Based Language Model, Interspeech, Vol. 2, No. 3

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis. *Journal of Knowledge-Based Systems*, Vol. 89, No. C, 14-46.

Ron, K. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 2, No. 12, 1137-1143.

Selvam, B. and Abiami, S. (2009). A survey on opinion mining framework. *International Journal of Advanced Research in Computer and Communication Engineering*, Vol.2, 3544-3549.

You, Q. (2016). Sentiment and emotion analysis for social multimedia: Methodologies and applications. *Proceedings of the 2016 ACM Multimedia Conference*, 1445-1449.

Vinodhini, G. and Chandrasekaran, R. M. (2012), Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, No. 6, 282-292.

(Received April 30, 2017; accepted June 5, 2017)