JAIST Repository
https://dspace.jaist.ac.jp/

| | |
|---|---|
| Title | レビューデータの意見分類に関する研究 |
| Author(s) | Ou, Wei |
| Citation | |
| Issue Date | 2018-12 |
| Type | Thesis or Dissertation |
| Text version | ETD |
| URL | http://hdl.handle.net/10119/15752 |
| Rights | |
| Description | Supervisor:Huynh Nam Van, 知識科学研究科, 博士 |

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

| | | | |
|---|---|---|---|
| 氏　　名 | OU Wei | | |
| 学位の種類 | 博士(知識科学) | | |
| 学位記番号 | 博知第 241 号 | | |
| 学位授与年月日 | 平成 30 年 12 月 21 日 | | |
| 論文題目 | A Study of Opinion Classification on Review Data | | |
| 論文審査委員 | 主査　HUYNH, Nam Van | 北陸先端科学技術大学院大学 | 准教授 |
| | 西本　一志 | 同 | 教授 |
| | 藤波　努 | 同 | 教授 |
| | DAM　HIEU CHI | 同 | 准教授 |
| | YAN Hongbin | East China Univ. of Sci. and Tech. | 教授 |

## 論文の内容の要旨

Nowadays, along with the rapid growth of the e-commerce economy, consumers have become more and more dependent upon review data to make decisions on shopping websites. However, reading reviews is a very time-consuming, even frustrating process when the number of reviews is overwhelmingly large. Both academia and industry have been working to devise algorithms that can automatically extract knowledge of products or services from the review data to improve users' review-reading experience. The knowledge extraction is usually treated as a sentence-level opinion classification problem, that is to detect what attributes of products or services consumers have discussed about, and how they felt of the attributes in the review sentences. This dissertation mainly focuses on the following 2 fundamental problems related to the classification task: sentence representations, the limited availability of training data. Also, based on the results of the opinion classification process, this dissertation proposes a novel machine learning based approach to identify the aspects that are generally attractive to consumers. The discovered attractive attributes allow users to quick capture the selling points of a product or service. A brief introduction to the originality of this dissertation is presented as follows.

1) Sentence representations.

Word embeddings models, as an effective way to represent text, have been widely used in various text classification tasks. Since word embeddings are only optimised to represent individual words, one has to define ways to aggregate word embeddings to represent sentences. A very effective, easy-to-compute aggregation function is averaging, though it obviously leads to loss of information. Recently, researchers have applied complex, but also computationally expensive neural network structures, such as convolutional neural network (CNN) and recursive neural network (RNN), to aggregate word embeddings. This dissertation proposes a novel weighted average approach, named `Abstract Keywords', as an alternative to the existing aggregation operators. The proposed approach assumes there exist some

extremely important abstract keywords that can be derived in the training process, and assigns words different weights according to their semantic similarities to the abstract keywords. Each sentence is represented by the weighed average of the embeddings of all words in the sentence. Experiment results show that the proposed approach is computationally efficient, and outperforms the simple averaging approach.

2) Limited availability of labelled training data.

As an important aspect of review mining, sentence-level sentiment classification has received much attention from both academia and industry. Many recently developed methods, especially the ones based on deep learning models, have centred around the task. Generally speaking, training sentence-level sentiment classifiers requires training datasets of labelled sentences, that are usually every expensive to obtain. It is possible to use the less expensive labelled review documents to train sentence-level sentiment classifiers, by treating each document as a long sentence, and the label of the document as the label for the long sentence. However, this way is obviously questionable because there may exist sentences in a document whose sentiments are very different from the sentiment of the document. Therefore, the sentiments of individual sentences can be easily misrepresented by the document-level labels in the training process. To address the problem, we propose a novel approach, named `Averaged-logits', that also uses labelled documents to train sentence-level sentiment classifiers, but makes a difference by assuming different sentences in a document have different sentiments, and the `average' of the sentence-level sentiments is used to determine the document-level sentiment. In the experiment, we collected two review datasets: one contains 50,000 hotel reviews crawled from TripAdvisor, the other 50,000 electronic product reviews from Amazon. The proposed approach was evaluated on the two datasets. The results show that, the proposed approach outperforms the existing approach treating each document as a long sentence, by margins of 3\%-8\% on sentence-level sentiment classification .

3) Attractive attribute classifiers.

Researchers have proposed statistical regression models that analyse on-line review data to identify attractive attributes of a product or service. This dissertation has the same aim, but with an approach based on machine learning models instead of statistical models. The proposed approach first extracts attribute-level sentiments from the review text by natural language processing techniques, then derives features that reflect the non-linear relations between attribute performance and customer satisfaction based on the sentiments. The non-linear features are fed to the Support Vector Machine (SVM) model to train predictive attractive attribute classifiers. The proposed approach is evaluated on a hotel review

dataset crawled from TripAdvisor. The experiment results indicate that the classifiers reach a precision of 79.3\% and outperform the existing statistical models by a margin of over 10\%.

**Keywords:** Review data; Sentiment analysis; Sentence representation; Weakly-supervised learning; Kano model analysis

## 論文審査の結果の要旨

Along with the rapid growth of e-commerce, opinion classification on review data has become an active research filed. The research of this dissertation focuses on the following three important problems related to the classification task: 1) the compositional representations of words, 2) limited availability of labelled training data, and 3) attractive attribute categorization. Three solutions have been respectively developed for these problems in this research as outlined below.

For the first problem, a novel weighted average model called ‘Abstract Keywords’ is proposed as an efficient alternative to the existing composition operators. This model first assigns different weights to different words in a review sentence based on the semantical similarities of the words to the abstract keyword(s), and then combines the embeddings of the words by their weighted average. Experimental results show that the proposed model has better performance than averaging, while it is more computationally efficient than CNN and RNN. For the second problem, this dissertation proposes a novel approach so-called ‘Averaged Logits’ that uses the prevalently available ratings as a form of document-level sentiment labels, instead of the manually annotated sentence-level sentiment labels, to train the sentence-level sentiment classifiers. Evaluation results show that, when the training size of the proposed approach is large enough, the performance of the proposed approach trained under the supervision of ratings can rival that of the classifiers trained under the supervision of the expensive sentence-level sentiment labels. As for the third problem, according to the Kano model theory, all the attributes of a product or service can be divided into 3 basic categories: must-be, attractive, one-dimensional. Automatically identifying the attractive attribute of a product or service from review data is particularly interesting as it allows on-line shoppers to instantly capture the ‘bright spot’ of a product or service. There exist two categories of methods that can exploit review data to identify the attractive attributes: penalty reward contrast analysis (PRCA), and critical incident technique (CIT). However, those existing methods cannot capture the high level of non-linearity in the asymmetric relations between attribute-level performance and customer overall satisfaction. In this dissertation, it is proposed to extract two types of discriminative features, called ‘empirical effect features’ and ‘interactive effect features’, to represent the asymmetric relations and train supervised attractive attribute classifiers based on the extract features. Evaluation results show that the proposed approach outperforms the

methods in PRCA and CIT by margins of over 10%.

This dissertation has made significant contributions to methodological and experimental developments within the area of sentiment analysis. The research work presented in this dissertation has resulted in two journal papers published, and two refereed conference papers.

In summary, Mr. OU Wei has completed all the requirements in the doctoral program of the School of Knowledge Science, JAIST and finished the examination on November 6, 2018, all committee members approved awarding her a doctoral degree in Knowledge Science.