Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes

Yoichi Saito*, Vitaly Klyuev*

*Software Engineering Lab, University of Aizu, Tsuruga, Ikki-machi, Aizu-wakamatsu City, Fukushima, 965-8580 Japan

m5211152@u-aizu.ac.jp, vkluev@u-aizu.ac.jp

Abstract— **Many products are sold on electronic commerce websites. Online customer reviews are available to help in selecting products to purchase. The products should be recommended by a special system that is capable to analyse and classify reviews because it is very hard for users to read many reviews and result of the recommendation should be personalized to suit user's requirements. The aim of this research is to classify the online customer reviews accurately to obtain opinion mining techniques of the recommendation system. The research focuses on classifying the Japanese reviews into positive or non-positive. In this study, we classify the reviews at the sentence and the review level. The data set for the sentence-level classification contains the reviews of Electronic Devices category. The data set for the review-level classification contains the reviews of Mobile Phone Accessories category. This research also compares the results of our experiments and another research to evaluate the experimental results. This research is successful to obtain opinion mining techniques and the better results at the review-level classifications instead of the sentence-level classifications. The experimental results will contribute to the opinion mining phase of the recommendation system.**

Keywords— **Sentiment Analysis, Opinion Mining, Naïve Bayes, Recommendation, Online Customer Reviews**

I. INTRODUCTION

Nowadays, recommendation systems have improved to promote customer activity on electronic commerce websites. On Amazon.co.jp, some related products are displayed recommended when the user inspects a product of his/her interest. Many methods are proposed to recommend the products to users. Traditionally, the main method to recommend products is collaborative filtering. Study [1] concluded that aspect-based opinion mining and collaborative filtering can recommend the products effectively. We cannot obtain the goodness of fit of each product for users if we utilize only opinion mining technique. We also cannot obtain the details of customers' opinion of each product if we utilize only collaborative filtering. Therefore, it is necessary to combine the results of opinion mining and collaborative filtering to obtain better results of recommendation.

Mining opinions from customer reviews is very important. The electronic commerce website stores a lot of customer reviews on each product. Users may read them to select the product to purchase, but it is hard for users to read all reviews on products in a category. So this labor needs to be reduced by utilizing opinion mining.

To solve this problem, we investigate opinion mining by analyzing the online customer reviews on the electronic commerce websites in this paper. We utilize a machine learning algorithm to classify the sentences into two categories: positive and non-positive. After that, we compare obtained accuracy with results of another research to evaluate the quality of our experiments. Results of this research can be used for opinion mining on the sentence level in the recommendation system.

II. RELATED WORK

One of the aims of opinion mining is to classify the documents into positive, negative or neutral. Generally, the documents are classified by machine learning algorithms in many studies. However, there are various methods of sentiment classification and accuracy evaluation depending on the researches. Moreover, the appropriate training data set does not exist in many cases. It is important to create appropriate data sets for the experiment and decide the appropriate method for opinion mining by considering the data properties and characteristics.

The performance of classification utilizing Naïve Bayes is compared with other machine learning algorithms in many cases because it is an effective method to classify the text data. The results show that the Naïve Bayes classification gives the better results on the IMDB, Twitter, Hotel, Amazon data sets compared to the Decision Tree and Support Vector Machine classification [2]. However, this research also mentions that it is effective to user machine learning algorithms if the length of documents is long, but it might not be effective to use machine learning algorithms if the length of document is short. In other words, it means that appropriate algorithms should be selected depending on the data set.

Another research proposes a new model to extract user and product level information for sentiment classification of documents [3]. This model is called User Product Neural Network (UPNN). The aim of this study is to investigate the effects of user and product information in the sentiment classification. As a result, the accuracy of classification by utilizing UPNN and user and product information is higher than the one by applying only UPNN and other several neural

network models. This research indicates the importance of aspect-based opinion mining by utilizing UPNN.

We need the Japanese data sets annotated with some polarities. However, the number of data sets created by third parties for Japanese sentiment analysis is extremely small. National Institute of Informatics [4] provides TSUKUBA Corpus as an example of Japanese annotated data sets. University of Tsukuba created this data set by annotating the review data on Rakuten Travel with positive or negative at the sentence level. This data set is not in the public domain. However, it is clear that it is inappropriate to utilize this data set for opinion mining on electronic commerce website because the words appear in the data set are different. Therefore, we created new data sets for our experiment.

III. DATA SETS

In this study, we collected Japanese online customer reviews of electronic commerce websites and analyzed the polarity of each review at the sentence level and the review level. The aim of this experiment is to classify the reviews into positive or non-positive. We created new data sets of customer review classification because there is not available data set on the Internet.

A. Data Collection

We collected the online customer reviews on products of Electronic Devices category from Amazon.co.jp and Mobile Phone Accessories category from Rakuten.co.jp. We utilized a free software Raku-review [5] to collect the reviews on Mobile Phone Accessories category. We collected the reviews on Electronic Devices category manually. The review data set of Electronic Devices is used for the sentence-level classification and the review data set of Mobile Phone Accessories is utilized for the review-level classification.

B. Data Preprocessing

In this section, we explain the preprocessing for the data set. We referred a study [6] for this step. We split all reviews of Electronic Devices category into sentences and completed removing the noises in order to properly format the sentences for classification. The character code of all characters in our data sets is UTF-8.

- Conversion of the all half-width Kana characters to the full-width Kana characters. Figure 1 shows an example. Hexadecimal code values for these symbols equal to ‘efbdb1’ and ‘e382a2’.

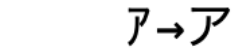

Figure 1. An example of corrections

- Conversion the all full-width digits and alphabets to the half-width characters. Figure 2 shows two examples. Hexadecimal code values for these symbols equal to ‘efbc90’, ‘30’, ‘efbd81’, and ‘61’.

Figure 2. Two examples of corrections

- Removal of all unnecessary symbols like ‘&’, ‘@’, ‘~’, ‘$’, ‘-‘, etc.
- Replacement of the all digits with ‘0’ and all alphabets with ‘X’. For example, a string like ‘Macbook’ is replaced with ‘X’ and ‘20180101’ is replaced with ‘0’.

It is impossible to remove the all digits and alphabets completely because it influences the result of the morphological analysis.

After removing the noise, we labelled sentences (reviews) as positive or non-positive manually without any techniques by ourselves. We created the relative frequency tables as the data sets from the labelled sentences (reviews) for classification by utilizing R software [7] and MeCab [8] and calculating the relative frequency for each word of every sentence (reviews). All stop words were removed at this step.

The formulas to calculated relative frequency of each word are defined as follows in R.

$$\frac{TF \cdot IDF}{\sqrt{\sum (TF \cdot IDF)^2}},$$
$$IDF = \log_2 \frac{N}{df} + 1,$$

where *TF* is row frequency for word, *N* is the number of all sentences (reviews), *df* is the number of sentences (reviews) the word appears. These formulas weight the words appearing in only certain sentences (reviews) instead of words appearing in every sentence (review). This formula also normalizes the correlative frequency because the length of every sentence (review) must be taken into consideration. ‘1’ is not added into *IDF* generally. However, the value of a logarithm is zero if the value of the anti-logarithm is 1 (In other words, the value of a logarithm is zero if the values of *N* and *df* are same) according to the mathematical definition. Even if the value of *TF* is not zero, the value of *TF* is not taken into consideration if the value of *IDF* is zero. ‘1’ is added into *IDF* in above formula to solve this issue [9].

The data sets created in these processes were divided into training data set and test data set in the experiments.

C. Statistical Information on the Data Sets

Table 1 shows the statistics on the discussed data sets. We utilized Japanese Sentiment Polarity Dictionary [10]–[12] (JSPD) to detect sentiment words in every sentences and reviews. A sentiment sentence includes sentiment words. There is no data of the number of reviews in the sentence-level data set because it was collected manually. “Number of Positive (Non-positives)” is the number of positive (non-positive) sentences (reviews).

TABLE 1. STATISTICS ON THE DATA SETS

Data Property	Electronic Devices (Sentence)	Mobile Phone Accessories (Review)
Number of All Words	49270	194778
Number of All Sentences	3253	12638
Number of Opinion Words	1886	8043
Number of Opinion Sentences	1459	5933
Number of Reviews	Not available	4904
Number of Positives	1759	2452
Number of Non-positives	1494	2452

IV. METHODS

In this study, we classify the reviews into positive or non-positive at the sentence level and the review level utilizing Naïve Bayes on the aforementioned data sets.

A. Naïve Bayes

Naïve Bayes [9] is one of the supervised learning classification methods. It is usually utilized for text classification in recent years because it can classify the sentences with the small amount of calculation in a high accuracy. Naïve Bayes classifies the sentences using conditional probability based on the naïve assumption of the independency and Bayes theorem. In this case, 'independency' means there are no relationships between the words. For example, it assumes the frequency of "good" and the one of "bad" do not influence each other. Actually, there may be relationships between frequencies of the aforementioned words, but the accuracy of the classification will be high and the amount of calculation will be small by ignoring the relationships.

Supervised classification methods of various kinds have been utilized for text classification. However, there are some good reasons to apply Naïve Bayes. Study [13] mentions the reasons to utilize Naïve Bayes to classify the documents. First, it provides competitive performance compared to the state-of-the-art discriminative classifiers such as Support Vector Machine, K-Nearest Neighbor, etc. because it is a model-based classification method with the naïve assumption of the independency. Second, it usually performs better at large vocabulary size when the document length and document class are assumed to be independent. Naïve Bayes is utilized for sentiment classification in this study for these reasons.

B. Data Separation

First, we created four pairs of data sets by splitting the data sets into 50% for the first training data set and 50% for the first test data set, 75% for the second training data set and 25% for the second test data set, 80% for the third training data set and 20% for the third test data set, and 90% for the last training data set and 10% for the last test data set of each classification level. The data sets of each pair must include the positive and non-positive sentences (reviews) at a persistent proportion. Second, we classified the reviews into positive or non-positive by utilizing the naiveBayes() classifier in R software [7] and the data sets of each pair at the sentence and the review level. Finally, we measured the accuracies of the classifications.

V. RESULTS AND DISCUSSIONS

A. Results

Table 2 shows the results of the experiment. Acc is one of the methods to measure the accuracy of classification [3]. It is calculated by dividing the number of sentences (reviews) classified correctly by the total number of sentences (reviews).

As a result, the highest accuracy at the review-level classification was obtained at both training and test data set size of 50%. The highest accuracy at the sentence-level classification was obtained at the training data set size of 75% and the test data set size of 25%. In the both classification level, changing the proportions of training and test data set did not improve the accuracy significantly. Overall, the accuracies at the review-level classifications are higher than at the sentence-level.

TABLE 2. ACCURACY OF CLASSIFICATIONS

Proportion	Electronic Devices (Sentence)	Mobile Phone Accessories (Review)
50%:50%	47.4%	50.5%
75%:25%	49.3%	50.2%
80%:20%	44.9%	50.3%
90%:10%	44.8%	50.1%

B. Discussions

First, we compared the results of the sentence-level and the review-level classification based on Table 2. The accuracies of the review-level classifications are higher than the sentence-level classification because the number of words per document at the review-level classification is larger than the sentence-level classification. If the number of words per document is small, the relative frequency table is extremely sparse. It is difficult to classify the reviews accurately based on the sparse table. Therefore, the accuracies of the sentence-level classifications are lower than the review-level classification because the relative frequency table is sparser than the one of the review-level classifications.

Second, we compared the results of the review-level classification and study [3] because the overall accuracy at the review-level classifications is higher than at the sentence-level classifications overall. This comparison is done on the basis of

publicly available evaluations. The author of study [3] utilized the same Naïve Bayes approach. Table 3 shows the data nature of utilized data sets. Table 4 shows the best classification result in our experiments and the accuracy obtained in study [3]. Acc represents the accuracy of classification (Higher is better), MAE represents Mean Absolute Error (Lower is better), and RMSE represents Root Mean Square Error (Lower is better). The result on IMDB data set is worse than the one on Yelp 2013, Yelp 2014 (English reviews), and our experiments. However, the accuracies of Yelp 2013 and Yelp 2014 data sets are higher by 10% than the best result accuracy of our experiments.

The rating scale of IMDB data set is 1-10 and the rating scale of Yelp 2013 and Yelp 2014 data set is 1-5 according to the research. In other words, the reviews of IMDB are classified into 10-scale ratings and the reviews of Yelp 2013 and Yelp 2014 are classified into 5-scale ratings in the experiment of the research. It is more difficult than our methods to classify the reviews accurately because the reviews are classified into only positive or non-positive at the review-level in our classification method. However, the reviews of Yelp 2013 and Yelp 2014 are classified more accurately than our review-level classification. It is because UPNN implements aspect-based opinion mining by utilizing the information of user and product accurately. However, it might be extremely difficult to classify the reviews on IMDB data set into 10-scale ratings accurately than our method even if the information of user and product on IMDB are utilized. Therefore, the accuracy of classification in our method is higher than the IMDB data set classification and lower than the Yelp 2013 and Yelp 2014 data sets.

TABLE 3. DATA NATURE

Data Set Name	Number of Reviews	Language	Domain
Mobile Phone Accessories	4904	Japanese	Product
IMDB	84919	English	Movie
Yelp 2013	78966	English	Stores such as restaurant
Yelp 2014	231163	English	

TABLE 4. COMPARISON OF CLASSIFICATIONS

Data Set Name	Acc	MAE	RMSE
Mobile Phone Accessories	0.505	0.495	0.703
IMDB	0.435	0.979	1.602
Yelp 2013	0.596	0.464	0.784
Yelp 2014	0.608	0.447	0.764

VI. CONCLUSIONS

In this paper, we classified the online customer reviews at the sentence and review levels by utilizing Naïve Bayes and compared the results with another research to evaluate our approach. The reviews of Electronic Devices category are utilized for the sentence-level classification. The reviews of Mobile Phone Accessories category are utilized for the review-level classification. We classified the reviews by incrementing the amount of training data set and decrementing the amount of test data set for each classification level. This research showed that changing the proportions of training and test data set does not improve the accuracy significantly and the accuracy at the review-level classification can be higher than at the sentence-level.

This research also showed that it is important for the data set to contain more reviews of various categories. To improve the accuracy of classification, it is necessary to prepare the larger size data set for training.

This study showed good results of the classification. We investigated the sentence-level and the review-level classifications. In our future work, we would like to combine opinion mining techniques obtained in this research with collaborative filtering to propose a user-oriented recommendation system.

REFERENCES

[1] Yao Wu and Martin Ester, "FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering" in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM'15)*, 2015, p. 119-208.

[2] Choi Y. and Lee H., *Data Properties and the performance of sentiment classification for electronic commerce applications*, Inf Syst Front (2017) 19: 993. https://doi.org/10.1007/s10796-017-9741-7. Springer, 2017, vol. 19.

[3] Duyu Tang, Bing Qin, and Ting Liu, "Learning Semantic Representations of Users and Products for Document Level Sentiment Classification" in *Proc. of the 53th Annual Meeting of the Association for Computational Linguistics and the 7th International Join Conference on Natural Language Processing*, 2015, p. 1014-1023.

[4] 楽天データ公開 | Rakuten Institute of Technology | 楽天技術研究所. [Online]. Available: http://rit.rakuten.co.jp/data_release_ja/

[5] らくれびゅ - スキマソフトウェア. [Online]. Available: http://soft.sukima.client.jp/rakurev/

[6] O. Abdelwahab, M. Bahgat, C. J. Lowrance, A. Elmaghraby, "Effect of Training Set Size on SVM and Naïve Bayes for Twitter Sentiment Analysis" in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015, p. 46-51.

[7] The Comprehensive R Archive Network. [Online]. Available: http://cran.ism.ac.jp

[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. [Online]. Available: http://taku910.github.io/mecab/

[9] Motohiro Ishida and Yuichiro Kobayashi, *Text Mining for Japanese Language with R*, 2nd ed, Tokyo, Japan: Hitsuji publisher, 2015.

[10] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, and Kenji Tateishi, "Collecting Evaluative Expressions for Opinion Extraction" *Journal of Natural Language Processing 12(3)*, pp. 203-222, 2005.

[11] Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto, "Learning Sentiment of Nouns from Selectional Preferences of Verbs and Adjectives" in *Proc. of the 14th Annual Meeting of the Association for Natural Language Processing*, 2008, p. 584-587.

[12] Jun Kikuchi and Vitaly Klyuev, "Gathering User Reviews for an Opinion Dictionary" in *Proc. of the 18th IEEE International Conference on Advanced Communications Technology (ICACT2016)*, February, 2016, Phoenix Park, Korea, p. 424-427, DOI: 10.1109/ICACT.2016.7423474.

[13] Paul M. Baggenstoss, Bo Tang, Haibo He, and Steven Kay, “A Baysian Classification Approach Using Class-Specific Features for Text Categorization” *IEEE Transactions on Knowledge and Data Engineering.*, vol. 28, pp. 1602–1606, June 2016.

Yoichi Saito is a master student in the Department of Computer Science and Engineering at the University of Aizu, Japan. His research areas are text and opinion mining.

Vitaly Klyuev is a professor at the University of Aizu, Japan. He received a Ph.D. degree in Physics and Mathematics from St. Petersburg State University, Russia in 1983 His research domain includes information retrieval, software engineering and analysis of computer algorithms. He has more than 100 publications in referred journals and conference proceedings, three co-authored and eight co-edited books. Dr. Klyuev is a member of editorial board of several academic journals and a program committee member of more than 20 conferences sponsored by ACM, FTRA, IEEE, ISCA and IARIA.