The Fourteenth IEEE International Conference on e-Business Engineering

A Rule-Based Chinese Sentiment Mining System with Self-Expanding Dictionary – Taking TripAdvisor as an Example

Jung-Bin Li
Dept. of Statistics and Information Management
Fu Jen Catholic University
New Taipei City, Taiwan
071635@mail.fju.edu.tw

Li-Bing Yang
Dept. of Statistics and Information Management
Fu Jen Catholic University
New Taipei City, Taiwan
libing0901@yahoo.com

***Abstract*—With the wide adoption of social networks, people are accustomed to post their ideas and thinking via these platforms. Tweets or comments online usually come with individual sentiment, which are time consuming to be analyzed by human labor. This study encapsulates a prototype Chinese sentiment mining system and takes a global hotel reviewing website TripAdvisor as the evaluation sample. The proposed sentiment mining model is compared with logistic regression and support vector machine models based on their performances. This proposed model outperforms LR and SVM in all datasets in terms of classification accuracy and F-measure.**

An additional module embedded in proposed system enables expansion of novel or undefined terms to the dictionary referred (NTUSD). With this Word2Vec-based module, the system further improves accuracy while reduces both type I and type II error for at least 5%.

Keywords-text mining, sentiment analysis, logistic regression, support vector machine, Word2Vec

I. INTRODUCTION *(HEADING 1)*

With the fast growing population of netizens, the aggregation of information on the Internet has become an important resource for businesses to explore their potential market or opportunities. However when the amount of data is too large to be process and analyzed by manpower, computer-assisted text mining tools enables businesses to do this job with efficiency.

Text mining is also applied in other fields, such as financial industry collects customer feedback opinions, interaction logs, investment diaries, web surfing records in to offer customized marketing experience. Shopping websites also take similar approach to push or recommend products.

Different from the case in western languages, applications of text mining in mandarin have to make a prior segmentation of document data into meaningful phrases, including semantic analysis, sentiment analysis, and automatic translation. Sentiment analysis identifies contents transformed by word segmentation, which is a straightforward task for manpower but is suffered when the data set is too large. With the evolution of languages, new words or phrases have to be constantly updated if the computerized sentiment system is adopted. Undefined sentiment phrases in the dictionary may influence the analysis outcome.

This study implements an automatic Mandarin sentiment analysis system in practice. The proposed system comprises segmentation module, sentiment analysis module, and phrase expansion module. It is evaluated by measuring the classification result given data collected from TripAdvisor website.

Sentiments can be classified by TF-IDF (Term Frequency – Inverse Document Frequency) [1] or semantic orientation [2]. This paper adopts the latter and implements an expansion module to collect new words or phrases not defined in existing sentiment dictionaries. The objectives of this study are:

- To examine necessary variables for sentiment analysis,
- To implement a sentiment mining system,
- To build an extra phrase expansion module, and
- To evaluate the proposed system.

II. PREVIOUS STUDIES

Text mining discovers valuable information from text databases. It generally refers to the retrieval of messages from unstructured text [3][4]. Tan [3] indicates that 80% of business data is in text form including documents, e-mails, memorandums, client mails and reports. Academic applications of text mining are information retrieval, topic tracking, data summary, classification, clustering, data visualization and response system [5]. Text mining is applied by public media to support decision-making processes [6]. Fields such business, health science, and political science also have text mining studies for exploring public data [7][8].

A. Sentiment mining

Sentiment analysis or opinion mining is a research field to discover the opinions, emotions, comments, or attitude of people to a certain target. Commodities, services, organizations or individuals, disputatious questions, or certain events are all potential targets in related studies [9]. Sentiment mining applies text mining mechanisms to analyze sentiment or opinions by software approaches.

Sentiment mining techniques include supervised machine learning, unsupervised machine learning, and unsupervised dictionary classification [10]. Supervised machine learning uses tagged data to predict untagged testing data, whose

978-1-5386-1412-9/17 $31.00 © 2017 IEEE
DOI 10.1109/ICEBE.2017.45

performance is decided by sufficient and reliable training data [10]. Its major algorithms include neural network, support vector machine, K's nearest neighbor, Naïve Bayes Classifier, and etc. [9][11]. Unsupervised machine learning, instead, uses untagged data as training source. Algorithms such as k-means clustering and hierarchical clustering belong to this category [11]. Dictionary affiliation builds sentiment phrase affiliations, and applies such relations to withdraw affiliated words [10].

Supervised machine learning has better performance than unsupervised model with the cost of time-consuming process for tagging and training data. Unsupervised learning in turn has the advantage of time efficiency [2].

Local sentiment mining studies focus on approaches of making classification rules. For sentiment mining in Mandarin, prior processing such as word segmentation is required. Studies mainly take advantage from open service providers like Yahoo! and Academia Sinica. Commonly referred sentiment dictionary is from cnki.net, NTUSD, and Chinese WordNet (CWN).

B. Sentiment computation

Previous studies classify comments positive or negative [9][12][13]. Some others classify certain emotions such as joy, anger, sorrow, and etc.[14] Fu and Wang distinguish emotions with intensity [15]. The classified emotions are referred by businesses to realize their attitude toward products or services.

Sentiments can be described with dynamic word and static word in Chinese. Static words express fixed sentiment, while dynamic words may vary with other phrases in sentence [15]. The but-clauses also change and even reverse the sentiment expressed in the sentence prior to it [16].

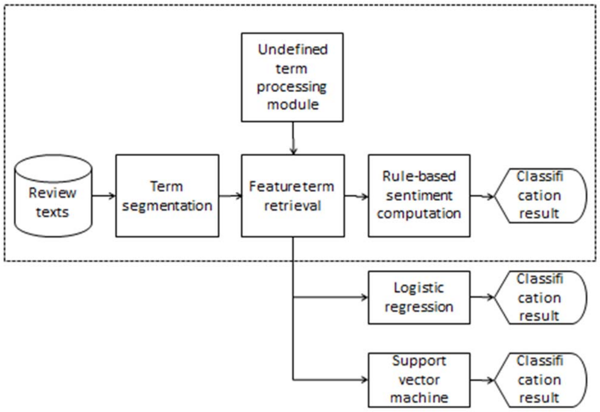


Figure 1. Framework of the proposed system

III. Proposed Sentiment Mining System

A. Framework

The framework of this proposed system (Figure 1) is a Software as a Service (SaaS) platform which is implemented by JAVA and encapsulated as an Application Programming Interface (API). Given comment text as system input, it provides positive / negative polarity output. For lexicons undefined in system dictionary, an embedded module assists in tagging the phrase and appends it into the dictionary.

To benchmark the performance of the proposed system, two popular classification models, logistic regression (LR) and support vector machine (SVM), are compared under the same condition.

This study takes the international hotel review site, TripAdvisor, as an example. The review posts of TripAdvisor are retrieved by web crawler and imported into the proposed system. The system proceeds with Chinese segmentation and computes the number of feature values. The positive / negative polarization output is based on the sentiment computation module. The LR and SVM module are given the same feature values to generate separate classification results for comparison.

B. Sentiment classification

This study classifies sentiment in positive, negative, and neutral polarization. For commentators who reveal joy and happiness in reviews, they are classified as positive. If anger or jealousy is found in reviews, commentators are classified as negative.

C. Chinese segmentation

Different from western languages, there is no separator such as space between phrases in Chinese. Hence text mining in Chinese has to segment articles into lexicons and phrases. An example of Chinese segmentation is as follows:

Before segmentation: 我跑馬拉松

After segmentation: [我][跑][馬拉松]

Chinese characters between brackets are separate lexicons. In this study, segmentation process is based on the Trie structure.

D. Sentiment analysis

Sentiment analysis classifies commentator's opinions by mechanisms such as Term Frequency – Inverse Document Frequency (TF-IDF) [1] or semantic orientation [2]. This study takes advantage of the latter and tags feature phrases from the sentiment dictionary proposed by [17].

A language modeling tool Word2Vec is also referred in the proposed system for transferring lexicons into vectors to compare the similarity among phrases [18][19]. Lexicons not in existing dictionary are processed to be classified as positive / negative and be appended in the dictionary.

Text sentiments are classified based on the variables retrieved after processes mentioned earlier, and compared with the given satisfaction scale. A confusion matrix is summarized for the evaluation of this system.

E. Feature value retrieval

A closer look of the text segmentation and feature value retrieval process is shown in Figure 2 in the next page.

Reviewer's comments are decomposed into review heading, review scale and the body text of review. By applying façade mode segmentation, lexicon dictionary can be expanded with specialized sets such medical, engineering, and other customized phrase set at any time.

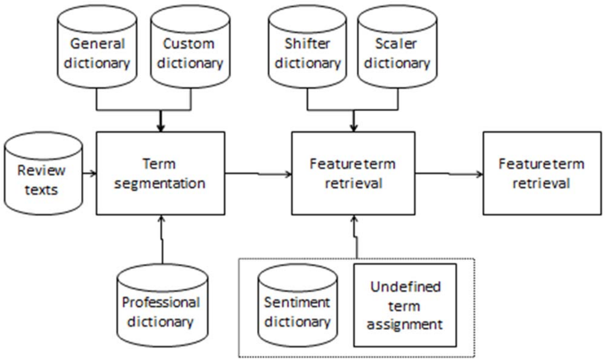
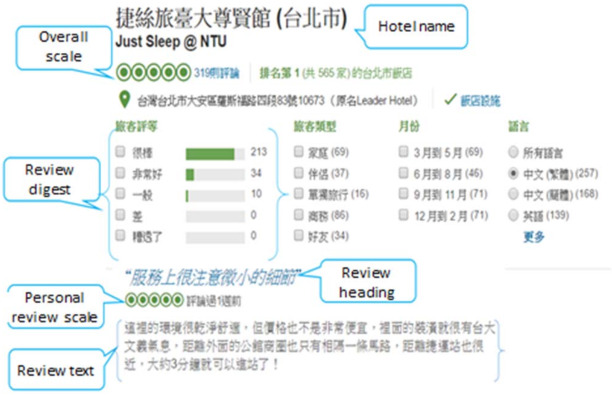


Figure 2. Text segmentation and feature term retrieval

Segmented lexicons are then processed by the feature phrase retrieval process. Lexicons listed in inverse phrase set, scale phrase set and sentiment phrase set are tagged and counted separately. The sums of each category, namely the feature values, are the input variables of the sentiment analysis module.

IV. Empirical Study – TripAdvisor

This study collects Chinese hotel review texts from TripAdvisor. A screen shot of TripAdvisor is shown in Figure 3. This study retrieves review texts by coding a web crawler. Since retrieved texts are wrapped with tags such as CSS, HTML, or JavaScript, a cleaning process has to be executed to collect clean information of hotel ID, body text of review, date, and overall recommendation scale.



捷絲旅臺大尊賢館 (台北市)
Just Sleep @ NTU
319則評論 | 排名第 1 (共 565 家) 的台北市飯店
台灣台北市大安區羅斯福路四段83號10673（原名Leader Hotel） ✓ 飯店設施

旅客評等		旅客類型	月份	語言
□ 很棒	213	□ 家庭 (69)	□ 3月到5月 (69)	○ 所有語言
□ 非常好	34	□ 伴侶 (37)	□ 6月到8月 (46)	● 中文 (繁體) (257)
□ 一般	10	□ 單獨旅行 (16)	□ 9月到 11月 (71)	○ 中文 (簡體) (168)
□ 差	0	□ 商務 (86)	□ 12月到 2月 (71)	○ 英語 (139)
□ 糟透了	0	□ 好友 (34)		更多

“服務上很注意微小的細節”
評論過1週前
這裡的環境很乾淨舒適，但價格也不是非常便宜，裡面的裝潢就很有台大文藝氣息，距離外面的公館商圈也只有相隔一條馬路，距離捷運站也很近，大約3分鐘就可以進站了！

Figure 3. Screen shot of TripAdvisor

A. Lexicon segmentation

The segmentation process in the proposed system applies an open source software package HanLP (https://github.com/hankcs/HanLP). The segmented phrases are further filtered to exclude unnecessary phrases, punctuation, and stop words.

B. Feature phrase retrieval

The variables adopted in this system are listed as follows:

- ANSWER: Sentiment classification of body text
- P_CNT: number of positive sentiment lexicons
- N_CNT: number of negative sentiment lexicons
- GET_ADV_CNT: number of scale lexicons
- GET_SHIFT_CNT: number of shifters / negations
- TOTAL_A_CNT: number of adjectives
- TOTAL_ADV_CNT: number of adverbs

ANSWER is the overall satisfaction scale given by the reviewer, from one (lowest) to five (highest). This study sets scale four and five as positive, three as neutral, and the remaining as negative. The overall scale is used as the reference output for model training of the proposed. Other variables are the model input.

C. Undefined sentiment lexicons

The sentiment dictionary adopted in this study is NTU Sentiment Dictionary (NTUSD). It has 8,276 negative terms and 2,810 positive terms. With the evolving nature of languages, the proposed system offers an expansion module which enables collection of new buzzwords. Expansion mechanism adopts Word2Vec and import the segmented phrase set to build a word vector model. By importing every term of sentiment dictionary stepwise, the first output term of adjective is appended into the dictionary.

D. Model evaluation

This study retrieves review comments from TripAdvisor. For over 95% of hotel reviews are positive (4/5 or 5/5 in overall satisfaction scale), training data of this study adopts stratified sampling to make the ratio of negatives and positives as 274/267, which is close to 1.

E. Tag and count of sentiment lexicons

The sentiment computation in the proposed system is illustrated in Figure 4.

If two sentiment lexicons are adjacent to each other, the combinations Positive-Negative (P-N) and N-P are usually regarded as negative. Such as the example below:

Original text: 墾丁大街的小吃種類[少]很[多]

Positive sentiment: [多]

Negative sentiment: [少]

A simple summation may neutralize the sentiment polarization. To better reflect the context, the computation module in this case disregards the positive count. For P-P or N-N cases, it is not necessary to make extra rules.

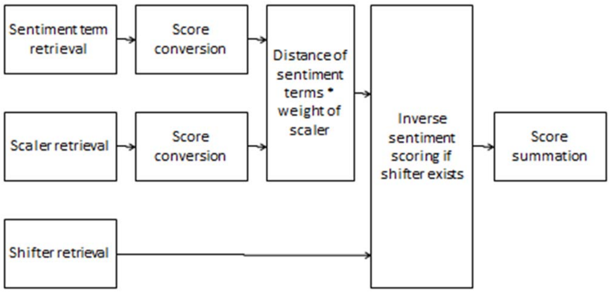


Figure 4. Model evaluation

F. Negation words

These words usually imply negative sentiments. Past studies indicate that sentences end with question marks to express doubt about certain fact can be classified as negative. This study regards interrogative sentences as shifters. An example is shown below. Negative lexicons are disregarded for transformation.

Original text: 沒學歷的你 要如何說服老闆你有[實力]能[勝任]此工作?

P-tag: 實力 勝任

N-tag: nil

P-tag after transformation: nil

N-tag: 實力 勝任

G. Sentiment scoring

This study counts positive, negative, and neutral term as +1, -1, and 0 point, respectively. The scoring process of the proposed is as follows:

Step 1: Extract sentiment lexicons, scale terms, and shifters (including question marks) from segmented text.

Step 2: Give each lexicon or term a score tag.

Step 3: Find scale terms whose distances from the sentiment lexicon are less than threshold value three, and multiply the scale term's mapping ratio provided from the scale term dictionary.

Step 4: Examine if there are shifters whose distances from the sentiment lexicon are less than threshold value three, or the sentence is ended with a question mark. In these two cases, transformation of lexicon scoring mentioned earlier is necessary.

Step 5: Compute the sentiment scoring by summarizing the scores received from Step 1 to Step 4.

H. Performance analysis

The confusion matrix of the proposed sentiment classification model is shown in Table I. The classification accuracy is 81.3%, and F-measure is 81.8%.

TABLE I. CONFUSION MATRIX OF PROPOSED MODEL

	Actual value		**Percentage**
	Positive	***Negative***	
Predicted positive	228	39	85.39%
Predicted negative	62	212	77.37%
Accuracy			81.3%

This study picks logistic regression (LR) and support vector machine (SVM) models for comparison. Both models are given the same data set cleaned and counted shown in Fig.1. The confusion matrices of LR and SVM are in Table II. F-measure of LR and SVM models are 83.3% and 87.1% respectively.

TABLE II. CONFUSION MATRIX OF LR AND SVM MODELS

	Actual value				**Percentage**	
	Positive		***Negative***			
	LR	SVM	LR	SVM	LR	SVM
Predicted positive	200	217	67	50	74.9%	81.3%
Predicted negative	62	37	212	237	77.4%	86.5%
Accuracy					76.2%	83.9%

To further evaluate three models, two additional data sets retrieved from reviews of hotels in Kaohsiung and Taichung city. A digest of data sets is shown in Table 3.

TABLE III. DIGEST OF ADDITIONAL TESTING DATA SETS

City	**# of P reviews**	**# of N reviews**	**% of P review**	**% of N review**
Kaohsiung	4718	158	96.75%	3.25%
Taichung	3253	137	95.95%	4.05%

The classification results of the reviews of two cities are summarized as in Table IV. For the case of Kaohsiung data set, the proposed sentiment classification model outperforms LR and SVM models, in terms of both accuracy and F-measure. The Taichung data set leads to a similar result where the performance of the proposed model is the best among the others.

TABLE IV. COMPARISON OF THREE MODELS

	Kaohsiung		**Taichung**	
	Accuracy	***F-measure***	***Accuracy***	***F-measure***
The proposed	91.8%	95.61%	88.71%	93.82%
LR	85.46%	91.93%	79.28%	88.04%
SVM	83.1%	90.51%	76.65%	86.31%

I. Sentiment dictionary expansion

The forementioned dictionary expansion module scans 24,724 review texts and compares the term vectors with the referred NTUSD. This process appends 128 positive and 57 negative sentiment terms. The updated sentiment dictionary is applied to model training process, and the performance analysis in the previous section is redone. The improvement

TABLE V. ACCURACY IMPROVEMENT WITH/WITHOUT DICTIONARY EXPANSION

	The proposed			**LR**			**SVM**		
	w/o	***w***	***Δ***	***w/o***	***w***	***Δ***	***w/o***	***w***	***Δ***
Kaohsiung	91.8%	93.1%	1.4%	85.5%	87.2%	1.9%	83.1%	84.7%	1.9%
Taichung	88.7%	90.6%	2.1%	79.3%	81.7%	3.1%	76.7%	78.9%	3.0%

TABLE VI. F-MEASURE IMPROVEMENT WITH/WITHOUT DICTIONARY EXPANSION

	The proposed			**LR**			**SVM**		
	w/o	***w***	***Δ***	***w/o***	***w***	***Δ***	***w/o***	***w***	***Δ***
Kaohsiung	95.6%	96.3%	0.7%	91.9%	92.9%	1.0%	90.5%	91.5%	1.1%
Taichung	93.8%	94.9%	1.2%	88.0%	89.6%	1.7%	86.3%	87.8%	1.7%

of each model is summarized in Table V and Table VI. The expansion of sentiment dictionary improves accuracy and F-measure of all three models. It implies periodical updates of dictionary are necessary, regardless of the classification model adopted.

V. CONCLUSIONS

Many approaches have been proposed in previous sentiment analysis studies, but few of them provide an efficient and systematic way to classify sentiment with evolving buzzwords in Chinese. This study integrates existing open source packages and proposes a sentiment mining system to find commentators' attitudes. The rule-based sentiment analysis module judges review texts by counting and calculating feature values.

Performance evaluation of the system is compared with logistic regression and support vector machine models, which are both popular for classification. For testing data sets collected from hotels in Kaohsiung and Taichung, our system has superior performance over LR and SVR models, with its advantage of rule-based nature of fast calculation. In addition, the dictionary expansion module improves the performance of every model. To further enhance the proposed system, texts of other varieties can be parsed and the weighing plan of sentiment strength can be optimized.

REFERENCES

[1] Joachims, T., "Text categorization with support vector machines: Learning with many relevant features", Machine learning: ECML-98: 137-142, 1998.

[2] Chaovalit, P., & Zhou, L., "Movie review mining: A comparison between supervised and unsupervised classification approaches", Proceedings of the 38th annual Hawaii international conference on system sciences, 2005.

[3] Tan, A.-H., "Text mining: promises and challenges", retrieved from https://www.researchgate.net/publication/2408427_Text_Mining_Promises_And_Challenges, 1999.

[4] Hearst, M. "What is TextMining?", Unpublished essay, 2003.

[5] Fan, W., Wallace, L., Rich, S., & Zhang, Z., "Tapping the power of text mining", Communications of the ACM, 49(9): 76-82, 2006.

[6] Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J., "An integrated text analytic framework for product defect discovery", Production and Operations Management, 24(6): 975-990, 2015.

[7] Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., & Gonzalez, G., "Utilizing social media data for pharmacovigilance: A review", Journal of biomedical informatics, 54: 202-212, 2015.

[8] Stieglitz, S., & Dang-Xuan, L., "Social media and political communication: a social media analytics framework", Social Network Analysis and Mining, 3(4): 1277-1291, 2013.

[9] Liu, B., "Sentiment analysis and opinion mining", Synthesis lectures on human language technologies, 5(1): 1-167, 2012.

[10] Muhammad, A., Wiratunga, N., & Lothian, R., "Contextual sentiment analysis for social media genres", Knowledge-Based Systems, 108: 92-101, 2016.

[11] Liu, B., "Web data mining: exploring hyperlinks, contents, and usage data", Springer Science & Business Media, 2007.

[12] Cambria, E., Schuller, B., Xia, Y., & Havasi, C., "New avenues in opinion mining and sentiment analysis", IEEE Intelligent Systems, 28(2): 15-21, 2013.

[13] Turney, P. D., "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", the Proceedings of the 40th annual meeting on association for computational linguistics, 2002.

[14] Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E., "Sentiment analysis: a review and comparative analysis of web services", Information Sciences, 311: 18-38, 2015.

[15] Fu, G., & Wang, X., "Chinese sentence-level sentiment classification based on fuzzy sets", the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010.

[16] Ding, X., Liu, B., & Yu, P. S., "A holistic lexicon-based approach to opinion mining", the Proceedings of the 2008 international conference on web search and data mining, 2008.

[17] Huang, T. H. K., Yu, H. C., & Chen, H. H., "Modeling pollyanna phenomena in Chinese sentiment analysis", the Proceedings of the 24th International Conf. on Computational Linguistics, Demo. Mumbai, India, pp. 231-238, 2012.

[18] Mikolov, T., Chen, K., Corrado, G., & Dean, J., "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, 2013.

[19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J., "Distributed representations of words and phrases and their compositionality", Advances in neural information processing systems, 2013.