International Journal of
Information
and
Management
Sciences

http://ijims.ms.tku.edu.tw/main.php

International Journal of
Information and Management Sciences

32 (2021), 101-119. DOI:10.6186/IJIMS.202106_32(2).0001

CODEN IIMSEQ

ISSN 1017-1819
DOI : 10.6186/IJIMS

International Journal
of
Information and Management Sciences

IJIMS http://ijims.ms.tku.edu.tw

101 Sentiment Analysis on Chinese Micro-blog Texts: A New Approach Using Enhanced Supervised Learning Model
Wei Shi, Huzhou University
Shaoyi He, California State University
Yue Fu, Huzhou University

121 A Hybrid Deep Learning Model for Predicting Stock Market Trend Prediction
Li-Chen Cheng, National Taipei University of Technology
Wen-Shiu Lin, Fu Jen Catholic University
Yu-Hsin Lien, Soochow University

141 Improvement of Firm Performance and Sustainability: Militancy, Strategic Entrepreneurship and Customer Focus, with Bank Agent Experience as Moderator
Osbal Saragi Rumahorbo, Sri Mangesti Rahayu, Suhadak, Solimun and Kadarisman Hidayat, Universiti Malaysia Pahang

157 Live-Streaming Consumer Experience; the Future of E-commerce and Digital Marketing in Africa: Evidence from China
Addo Prince Clement, Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development, and University of Electronic Science and Technology of China
Kulbo Nora Bakabbey and Ahrong Godwin Banafo, University of Electronic Science and Technology of China
Asare Ohemeng Andy, George Brown College

177 Donor Experience Sharing, Why, and Who They Are?
Horng Jinh Chang and Ming Han Du, Tamkang University

Volume 32, Number 2 June, 2021

Information and Management Sciences

June, 2021 Volume 32, Number 2

Sentiment Analysis on Chinese Micro-blog Texts: A New Approach Using Enhanced Supervised Learning Model

Wei Shi[1], *Shaoyi He*[2] *and Yue Fu*[1]
[1]Huzhou University and [2]California State University

Keywords

Sentiment analysis
Micro-blog text
Opinion mining
Enhanced supervised learning

Abstract.

The sentiment analysis of the review texts of micro-blog is helpful for deep mining Chinese Micro-blog (Weibo) - one of the main social media in China. Aiming at the shortcomings of the widely used machine language in sentiment analysis of texts when dealing with sentences containing connectives, this paper formulates rules for dealing with Chinese connectives, incorporates expression symbols into feature vectors, calculates sentiment decision scores with sentiment dictionaries, and proposes an enhanced supervised learning model that is based on language rules and emotional scores. Examples show that the proposed model can significantly improve the effectiveness of text classification.

1. Introduction

Along with the development of social media technologies, more and more people have been using social media for different situations in their daily life. For example, people use social media to express their opinions on different topics (see Cambria, et al. [13]; Neubaum and Kr`‘amer [40]; Bergstr`‘om and Jervelycke [9]; Baum and Potter [8]). People also use social media for many other situations, such as: in B2B communication to improve business performance (see Wang, Pauleen and Zhang [58]), in disaster planning, response, and research (see Houston, et al. [27]), in marketing academic library information services (see AlAwadhi and Al-Daihani [4]), in measuring user satisfaction (see Balbi, Misuraca and Scepi [6]), and in emotion and sentiment detection, classification as a special type of social media and analysis (see Chaturvedi et al. [15]; Rout et al. [45]; Tang et al. [51]). Specifically, micro-blogging, a special type of social media represented by Twitter in the USA and Micro-blog in China, has been used extensively for emotion and sentiment detection, classification and analysis (see Abudalfa and Ahmed [2]; Cureg et al. [19]; Hasan, Rundensteiner and Agu [25]; López [36]; Kaur, Pannu and Malhi [31]; Mostafa [38]; Sailunaz and Alhajj [47]; Tyagi et al. [54]). Since micro-blog texts are different from other ordinary texts such as web news, blog articles and so on, their characteristics have been identified and analyzed by scholars and researchers in the past years (see Java et al. [28]; Ellen[21]; Zhang et al. [65]; Yuan and Purver [63]; Cheng et al. [16]; Wu and Wang [60]; He et al. [26]). According to them, micro-blog texts have some main characteristics as follows:

(1) They are short in length. A micro-blog text generally cannot exceed 140 words, many micro-blog texts usually only have one or two sentences or even a few words or phrases;
(2) The grammar is not standardized. Unlike webpage news or blog articles, micro-blog texts are usually posted without careful consideration, so the content is very arbitrary, contains many erroneous words, spoken language, abbreviations, emoticons, hyperlinks and other noises;
(3) The content is highly contextual. Many micro-blog texts are replies to other posts, agreeing or disagreeing with them, commenting or evaluating them, or simply expressing opinions.
(4) The number of micro-blogs is huge. Because of the low threshold of micro-blogs, anyone can publish any information with a micro-blog, so the speed of information publishing is unparalleled by any other media, which makes people quickly submerged in a huge amount of micro-blog text information;
(5) They contain a lot of valuable hidden information that is emotion-based and/or opinion-related.

The above characteristics make classifying and extracting emotions from micro-blog texts extremely challenging. There are many ways to classify and extract emotions in micro-blog texts. We can calculate the emotional class of text based on sentiment dictionary/lexicon and semantic analysis (see Yao and Lou [61]; Medhat, Hassan and Korashy [37]; Yuan et al. [62]; Chopade [18]; Ai et al. [3]; Zhang et al. [66]; Kamal et al. [30]; Li et al. [32]; Pan, Mou and Liu [41]). We can also conduct effective data mining and sentiment analysis of review texts based on machine learning (see Pang and Lee [42]; Liu and Liu [35]; Jiang et al. [29]; Hasan et al. [24]; Sohangir et al. [50]; Zhang, Wang and Liu [66]; Ramanathan and Meyyappan [44]; Vo and Ngoc [55]). This paper focuses on sentiment analysis on the microblogging texts collected from Chinese microblogs, because they contain a large number of emotional reviews that have great value in providing useful and helpful insights for a better understanding of the situation via the texts (Shi, Wang and He [48]; Lin et el. [33]; Zhang et al. [66]; Geng et al. [22]; Wang et el.[57]). Traditional sentiment analysis has some limitations, sentiment analysis based on sentiment dictionary is limited by the quality and coverage of sentiment dictionary and semantic rules. Many researches on sentiment analysis of Chinese microblog do not consider the influence of conjunctions and emoticons on semantic sentiment. In sentiment analysis based on machine learning, model training depends on the quality of annotated data sets, and high quality data sets require a lot of labor costs. In view of these characteristics of micro-blog texts and the shortcomings of traditional sentiment analysis methods, this paper proposes a new method of sentiment analysis on micro-blog texts using an enhanced supervised learning model. The main contribution of this paper is to classify the micro-blog texts according to their polarity level, including predicting the sentiment classes in the text, and adopting new methods to improve the effectiveness of the sentiments classification in micro-blog texts.

The remainder of this paper is organized as follows: Section 2 provides a review of relevant literature and explains the motivation for our research. Section 3 puts forward research ideas and methods. Section 4 carries out relevant experiments and analyzes the results. Section 5 provides a summary this paper and points out the future research directions.

2. Literature Review

In this section, we first outline the literatures related to the research methods of this paper, and then propose the motivation of this study through literatures and experiments.

2.1. Sentiment analysis on micro-blog texts

Sentiment analysis on micro-blog texts was first carried out on Twitter (see Balijepalli [7]). Later on, some studies have conducted sentiment analysis of Twitter public publications, using the extended version of POMS (Profile Of Mood States) to extract emotions, comparing the results of sentiment analysis with the fluctuations of major events in the stock market, crude oil price indicators and the media at the same time, and finding events occurring in the social, political, cultural and economic fields. It has a distinct immediate impact on different aspects of public sentiment (see Go, Bhayani and Huang [23]; Bollen et al. [10]; Agarwal et al. [5]; Bollen, Mao and Pepe [10]; Bollen, Mao and Zeng [12]). After 2012, sentiment analysis on Chinese social short texts represented by Chinese micro-blog emerged one after another. Liu [34] used three machine learning algorithms, three feature selection algorithms and three feature weight calculation methods to conduct an empirical research on emotional classification of Chinese micro-blog texts, and achieved good results. Ding et al. [20] used CRFs model to identify viewpoint sentences in Chinese micro-blog. On the premise of ensuring accuracy, the recall rate was raised to 61.8represent micro-blog texts, and proposed a method of micro-blog text orientation analysis which combines pattern matching and machine learning. The experimental results show that this method is effective compared with other evaluation results. Zhou et al. [67] proposed a neural network model with multi-window and multi-pool layers, which effectively utilized semantic dependency distance and multi-level semantic to classify Chinese micro-blog texts emotionally. The model was validated by Stanford Emotion Tree Database dataset and achieved good results. Li et al. [32] proposed an automatic building method of emotion lexicon based on the psychological theory of compound emotion, which could map emotional words into an emotion space, and annotate different emotion classes through a cascade clustering algorithm. The experimental results show that their method outperforms the state-of-the-art methods in primary classification performance on both word and sentence-level, and also offer some insights into compound emotion analysis.

2.2. Supervised learning and unsupervised learning for sentiment analysis

Supervised learning infers a functional machine learning task from marked training data. A text sample is converted to a feature vector that represents its most important feature. The most common feature of sentiment analysis is the existence of a single tag or uniform symbol term and the term frequency. Wiegand et al. [59] focus on expressing negative techniques, detecting negative words, determining the scope of negation in the text, and achieving good results. Chikersal et al. [17] enhance sentiment analysis on Twitter with micro-blog text or Twitter-specific features such as emoji, tags, URLs, @symbols, uppercase letters, and extended words. The corresponding research goals have been achieved. Rahimi, Noferesti and Shansfard [43] implement a supervised machine learning system based on semantic features of sentences for shifter identification and polarity classification, they test their proposed algorithms on polarity classification task for 2 domains: a specific domain and a general domain, the proposed semantic based machine learning method performs well in polarity classification.

In machine learning, the problem of unsupervised learning is trying to find hidden structures in unlabeled data. Turney [53] first creates an emotional vocabulary in an unsupervised way, and then uses a function to determine the polarity of the text based on the number or metric of positive and negative words and/or phrases that appear in the text. Chaovalit and Zhou [14] conduct a comparative study on supervision methods and some unsupervised methods, and analyze their respective advantages and disadvantages. Abid et al. [1] construct a joint architecture by first placing RNN first, using the global average pool to capture long-term dependencies with CNN, and then using unsupervised learning to process words based on a large number of Twitter corpora.

2.3. Experiments and research questions

This research trains Support Vector Machine (SVM) classifiers with about 10,000 Chinese micro-blog texts, and tests them with N-gram language model. The confusion matrix function calculated by SVM in Figure 1 is drawn as follows:

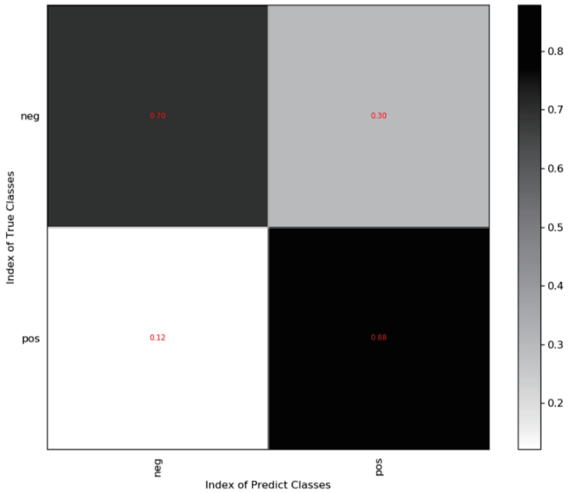


Figure 1: Confusion Matrix Function Diagram of SVM Classification Results.

In Figure 1, the black squares indicate correct classification, white and grey squares indicate wrong classification, and small values represent proportion. Further querying the decision scores of misclassified texts, we find that most of them have decision scores ranging from -0.5 to $+0.5$, such scores mean that SVMs are not sure about them, which leads to misclassification.

Table 1: Result Report for Confusion Matrix of SVM Classification Results.

	precision	recall	f1-score
neg	0.66	0.70	0.68
pos	0.90	0.88	0.89
avg / total	0.84	0.83	0.84
confusion_matrix	[[701 300], [365 2635]]		
acc_for_each_class	[0.66 0.9]		
average_accuracy	0.777692		
overall_accuracy	0.833792		
score	0.833792		

Table 1 shows the specific data results of the confusion matrix of SVM classification results. Precision represents the accuracy of sentiment orientation classification of micro-blog texts, recall represents the recall rate of sentiment orientation classification of micro-blog texts, and f1-score is the weighted average of accuracy and recall rate. About 66positive sentiment can be accurately classified, and about 90classified. However, judging from the calculation results of confusion matrix, most of the texts can not be classified or classified incorrectly.

Although many efforts are being spent on sentiment analysis for Chinese Micro-blog Texts, current studies still have limitations: (1) Supervised Learning Classifier for Polarity Classification relies on feature vectors extracted from text to represent the most important features of text, it is not effective enough in dealing with sentences with sentence patterns (see Tripathy, Agrawal and Rath [52]; Zhou et al. [68]).(2)When modal verbs such as “可能” (possible), “應該” (ought), “也許” (maybe) or conjunctions such as “但是” (but), “既然” (since), “如果” (if) and so on appear in Chinese sentences, they will greatly increase the difficulty of forecasting supervisory classification. (3) Many people often use sentiment words or emoticons repeatedly, and the addition of emoticons will greatly affect the sentiment tendency of the review text in many times, there is a lack of sentiment rules of emoticons in Chinese micro-blog reviews.

Based on literature review and experiments, we pose the following research questions:

(1) How to deal with special grammatical parts of sentences, such as conjunctions?

(2) How to verify or change the classification label of micro-blog text with low decision score calculated by support vector machine? How to deal with the special part of grammar?

(3) What is an emoticon dictionary? How to make emoticons sentiment rules in Chinese microblog sentiment analysis?

3. Research Design

3.1. Data preprocessing

Before analyzing the sentiment of the original micro-blog texts, we preprocessed them. During the preprocessing, all @<username> references were changed to @USER, and all URLs are changed to http://URL.com. Then, we extracted and assigned some tags to the micro-blog texts. In addition to nouns, verbs, adjectives and adverbs, this method can also assign tags to conjunctions and some specific contents in micro-blog texts such as emoticons and URLs. The sentiment analysis system designed in this paper is shown in Figure 2.

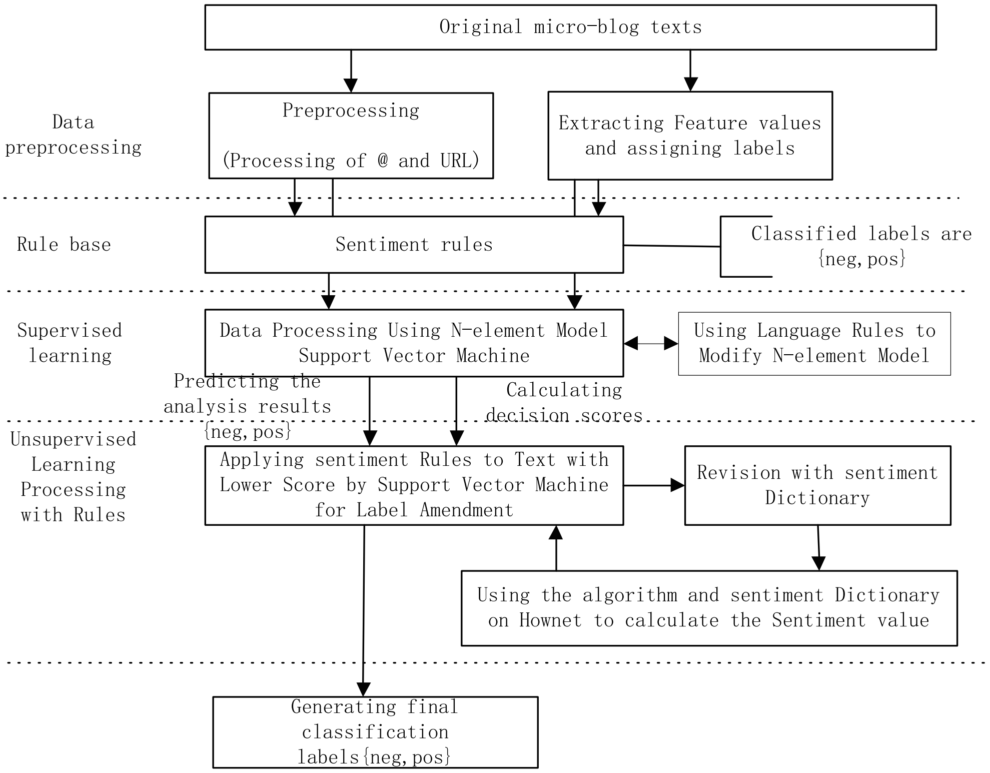
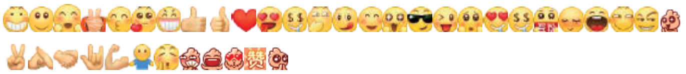
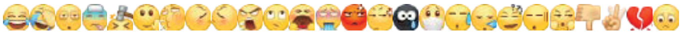

Original micro-blog texts
Data preprocessing
Preprocessing
(Processing of @ and URL)
Extracting Feature values and assigning labels
Rule base
Sentiment rules
Classified labels are {neg, pos}
Supervised learning
Data Processing Using N-element Model Support Vector Machine
Using Language Rules to Modify N-element Model
Predicting the analysis results {neg, pos}
Calculating decision scores
Unsupervised Learning Processing with Rules
Applying sentiment Rules to Text with Lower Score by Support Vector Machine for Label Amendment
Revision with sentiment Dictionary
Using the algorithm and sentiment Dictionary on Hownet to calculate the Sentiment value
Generating final classification labels{neg, pos}

Figure 2: Flow Chart of a Sentiment Analysis System Based on Enhanced Supervised Learning.

3.2. Sentiment rule-making

The traditional method of texts classification by extracting feature values of texts has a high accuracy in most cases, but it will encounter problems when processing micro-blog review texts. In order to emphasize their opinions, many people often use emotional words or emoticons repeatedly in Chinese micro-blog reviews, and the addition of emoticons will greatly affect the sentiment tendency of the review text in many times (see Shi, Wang and He [49]). To this end, a dictionary of emoticons is compiled to determine the sentiment tendency of sentences by counting positive and negative emoticons.

Table 2: Manual Dictionary of Emoticons in Chinese Micro-blog.

Sentiment class	Emoticons
Positive	
Negative	

Sentiment rules of emoticons are as follows:

(1) If a micro-blog text contains one or more positive emoticons without negative emoticons, it is marked as positive.

(2) If a micro-blog text contains one or more negative emoticons without positive emoticons, it is marked as negative.

(3) If neither of the above two rules is applicable, the micro-blog text is marked as unknown.

Using these sentiment rules, we can judge the sentiment class of micro-blog text and add sentiment labels to the text. The micro-blog texts marked as unknown with this rule are transmitted to the next stage of the sentiment analysis system–supervised learning classifier.

3.3. Supervised learning using N-gram language model of Support Vector Machine (SVM)

N-gram is a language model commonly used in large vocabulary continuous speech recognition. For Chinese, it is usually called Chinese Language Model (CLM). Using the collocation information between adjacent words in the context, the Chinese language model can compute the sentences with the most probable rate when it is necessary to convert the continuous blank-free Pinyin, strokes or numerals representing letters or strokes into Chinese character strings (i.e., sentences), so as to realize the automatic conversion of Chinese characters without the need for user's manual selection. It avoids the problem that many Chinese characters correspond to the same Pinyin (or strokes or numbers).

This paper adopts the following text processing method: adding string "_NEG" after all negative words extracted from micro-blog text to illustrate the negative. All nouns, adjectives, adverbs or verbs extracted from the micro-blog text between the negative word and the next punctuation mark are considered negative. In addition, other features related to negation are not used in feature vectors. After processing the micro-blog text with the above methods, the SVM algorithm of sklearner module in Python language is used to supervise learning and add sentiment classification labels.

3.4. Modify N-gram language model according to language rules

Typical supervised learning methods based on N-gram language model can obtain satisfactory results for text sentences without conjunctions, but problems arise when processing text with special sentence structures such as conjunctions. Using the Chinese grammar rules, this paper manually determines analysis methods including conjunctions such as “但” (but), “可是” (however), “如果” (if), “除非” (unless) , “萬一” (just in case)and so on, and formulates supervisory learning rules that can remove irrelevant or opposite sentiment tendencies from text feature vectors. There are many uses of conjunctions in Chinese, among which the adversarial conjunction “但” (but) and hypothetical conjunction “如果” (if) are the most widely used. This paper focuses on the analysis of the impact of conjunctions represented by “但” (but) and “如果” (if) on the sentiment analysis of Chinese Micro-blog texts.

3.4.1. Text Processing Strategy with the Conjunction “但” (but)

Table 3 lists several examples of text with “但” (but) in different grammatical positions, and we determine the overall polarity of the text manually. From the following examples, it can be seen that part of the sentence after “但” (but) in examples I to III usually shows the overall polarity of the text better than the previous part. In example IV, it is difficult to determine the most important part of the text. Because this microblog text is only weak affirmative, if we only consider “但” (but) in the second part, it can even be judged as negative. In Example V, it is difficult to determine which part is emphasized by the user. Processing text similar to IV and V is too difficult and requires more complex language rules. In this paper, we focus on the rule of text sentiment classification similar to example sentences I to III, and deal with the text similar to IV and V using to this rule.

Therefore, the following strategies are proposed to modify the text containing the conjunction “但” (but) (similar to the text containing the conjunction “但是” (but) and “可是” (but):

(1) Firstly, the text is segmented by using the Jieba module in Python language, and the stop words are removed.

(2) In each sentence, find the last position where “但” (but) appears.

(3) Delete all the characters before this position, so that the adjusted sentence contains only the characters after the last “but“ position.

(4) Once all the sentences in the text are processed, the revised sentences are merged to get the revised text.

3.4.2. Text processing strategy with the conjunction “如果” (if)

Table 4 lists several examples of “如果” (if) text with different grammatical positions, and determines the overall polarity of the text manually.

Table 3: Example Sentences Containing “但” (but) and Their Polarity.

Example Sentences	Polarity
I. 今天過的很糟糕, **但**一想到我們可以明天一起在必勝客吃飯就開心了。(It’s been a bad day, **but** I’m happy when thinking about having dinner at Pizza Hut tomorrow.)	Positive
II. 面試失敗, **但**說不定明天去的那家更好。(The interview failed, **but** the one to go tomorrow maybe better.)	Positive
III. 你輸了比賽却贏了我的心你可能不認識我, **但**我崇拜你。(You lost the game but won my heart. You may not know me, **but** I admire you.)	Positive
IV. 我覺得這部电影很好看, **但**没有和任何人分享。明天天氣會變好嗎？（I thought it was a good movie, **but** I didn’t share it with anyone. Will the weather be better tomorrow?)	Positive
V. 天啊, 我錯過了星期六的比賽! **但**是有更重要的事情等着我去做。(My God, I missed Saturday’s game! **But** there are more important things waiting for me.	Negative

Table 4: Text Example Sentences Containing “如果” (if) and Their Polarity.

Example Sentences	Polarity
I. 如果北京國安在中超聯賽中没有取得最高積分, 那麼他肯定是第二强的!(**If** Beijing Guoan did not get the highest score in the Chinese Super League, then it must be the second best!)	Positive
II. 如果明天你不去现场看比赛的話, 可以在电视上看直播。(**If** you don’t go to the game tomorrow, you can watch it live on TV.)	Positive
III. 如果姚明還在火箭隊的話我會看直播的 ··· 我不喜歡火箭隊但我喜歡他。(English: **If** Yao Ming is still in the Rockets, I will watch the live broadcasting .... I don’t like the Rockets, but I like him.)	Positive
IV. 如果你也是趙麗颖粉絲的話我挺你。(English: **If** you’re a fan of Zhao Liying, I’ll stand by you.)	Positive
V. 如果你明天不去參加下午5:30開始的訓練, 你將失去比賽資格, 希望你能來!(English: **If** you don’t go to the training starting at 5:30 p.m. tomorrow, you will lose your qualification. Hope you can come!)	Negative
VI. @USER能來參加十月份的北京錦標賽嗎？我一直想見到你, 如果你在這里參賽的話會很令人激動! (English: @USER Can you come to Beijing Tournament in October? I want to see you as always, **if** you are here , it would be very exciting!)	Positive

From the examples above, we can see that “如果” (if) has more syntactic positions than “但” (but), we can specify the following conditions:

(1) If <conditional clause> then <result clause>

(2) If <conditional clause>, <result clause>

(3) If <conditional clause> <ellipsis then / comma or other> <result clause>

(4) <result clause> If <condition clause>

According to grammatical rules, case I belongs to condition (1), case II, III and V belong to condition (2), case IV belongs to type (3), and case VI belongs to type (4). In the examples I and II, the most important part of the text is the part that appears after the comma. In Example III, the most important part is the part after the first comma. However, example III contains both “如果” (if) and “但” (but), which makes it more difficult to automatically determine the strongest part. In addition, in examples IV and V, due to grammatical errors and informality of the text, the most important part is not after the comma. In example VI, “如果” (if) appears in the middle of a sentence, it is difficult to automatically determine the range of the strongest part. Determining the most important parts of sentences like IV, V, and VI requires more complex linguistic analysis, which goes beyond the scope of this study.

Therefore, the following strategies are proposed to modify the N-gram language model of the text containing the conditional sentence “如果” (if) which is similar to the text containing “除非” (unless) and “萬一” (in case):

(1) Firstly, the text is segmented using the Jieba module in Python language, and the stop words are removed.

(2) In each sentence, find the last position of the conditional sentence.

(3) Find the first comma after the conditional clause and the position of “如果” (if).

(4) Delete all characters between conditional clauses and commas, and delete conditional clauses and commas. The remaining parts make up the revised sentences.

(5) Once all the sentences in the text are processed, the revised sentences are merged to get the revised text. If a text contains both “如果” (if) and the conjunction“但” (but), only the “但” (but) rule is applied.

Finally, for each improved text, a modified eigenvalue vector (the text eigenvalue vector is modified according to the above five rules) is created. The N-gram language model is improved by using language rules, and then processed by Support Vector Machine (SVM) and added classification labels.

With the linguistic rules to improve the N-gram language model, the steps of calculating the sentiment tendency score of micro-blog text are shown in algorithm 1.

Algorithm 1. Computing Pseudo-code for Sentiment Tendency Score of Micro-blog Text

Input: Read degree adverb dictionary, stop word dictionary and affective word dictionary from database

Output: Sentiment Tendency Score of Micro-blog Text

```
Segment a sentence to words
Delete stopwords
return newSent
for word in newSent:
token1 = [ ]; pos1 = [ ];
for each token, tag in tokens, pos do
    if token is NOT a stopword then
        append token to token1 and tag to pos1
    end if
end for
concepts = [ ]
conceptTagPairs = [(“N”, “N”), (“N”, “V”), (“V”, “N”), (“A”, “N”), (“R”, “N”), (“P”, “N”),(“P”, “V”)] # “N” = Noun, “V” = Verb, “A” = Adjective, “R” = Adverb, “P” = Preposition
for ti in range(0, len(tokens1)) do
    token = tokens1[ti]; tag = pos1[ti];
    prevtoken = tokens1[ti-1]; prevtag = pos1[ti-1];
    token_stem = Stem(token); prevtoken_stem = Stem(prevtoken);
    {raw tokens and stemmed tokens are extracted as single-word concepts}
    append token to concepts
    append token_stem to concepts
    if (prevtag, tag) in conceptTagPairs then
        append prevtoken+“ ”+token to concepts
        append prevtoken_stem+“ ”+token_stem to concepts
    end if
end for
extract the sentiScore from the Emotional dictionary
finalSentiScore = (−1)^(num of notWords) * degreeNum * sentiScore
finalScore = sum(finalSentiScore)
```

3.5. Adjusted Support Vector Machine (SVM) prediction results

In the training process, support vector machine (SVM) approaches the optimal decision boundary of a separated data point (sample eigenvector) belonging to N different classifications ($N = 2$, classification positive, negative). The data points supporting the decision boundary are called support vectors. Each trained SVM has a scoring function, which calculates the decision score of each new sample and assigns the classification label based on it. The scoring range of SVM decision-making for sample classification is from sample eigenvector x to decision-making boundary, which is calculated by the following formula:

$$\text{SVM Decision Score} = \sum_{i=1}^{m} \alpha_i y_i G(x_i, x) + b. \tag{3.1}$$

Among them, $\alpha_1, \alpha_2, \ldots, \alpha_m$ and b are estimated by SVM, $G(x_i, x)$ is the point product in the prediction space between x and SVM, and m is the number of training samples.

As mentioned in Part 2, when using N-gram model, the decision scores of a large number of texts are too low, this is because their eigenvectors are very close to the decision boundary, support vector machines are not sure what labels to assign to them. Therefore, after supervising and classifying all unmarked texts, the reliability of SVM prediction is determined according to the decision score derived from each decision. For texts whose decision scores are close to decision boundaries or whose confidence is less than 0.5, class labels allocated by SVMs are discarded and unsupervised classification method is used to predict their classification labels. The unsupervised classification process is as follows:

(1) In order to take into account conjunctions and conditional clauses, the method described in Part 3 is used to modify the text.

(2) Use algorithm to extract keywords from text.

(3) Query all these sentiment words in HowNet and calculate the sentiment value (see Wei Shi et al. [49]). The calculation method is as follows: degree words as weight multiplied by sentiment word score (We extract 60 degree words from the HowNet and divide them into 7 classes, the setting is listed in Table 5), negative words multiplied by weight -1, normalized processing was used as the sentiment score of the sentence.

(4) Find the number and sentiment value of positive and negative conceptual words:

Table 5: Assignment value of degree words (see Wei Shi et al. [48]).

Value	Degree Words
1.5	最 (bottom)、最爲 (most)、極 (mighty)、極爲 (very)、極其 (spanking)、極度 (to the utmost)、極端 (exceeding)
1.4	太 (so much)、绝 (absolutely)、至爲 (to the)、顶 (top)、(over)、過于 (excessively)、過分 (overmuch)、分 (exceptionally)、万分 (extremely)、何等 (how)
1.3	很 (quite)、挺 (rather)、怪 (odd)、老 (always)、非常 (greatly)、特别 (special)、相當 (quite)、十分 (great)、甚 (very)、甚爲 (very)、异常 (remarkably)、深爲 (deeply)、蛮 (pretty)、滿 (completely)、够 (really)、多 (much)、多麼 (so)、殊 (outstanding)、何其 (how)、尤其 (especially)、無比 (unequaled)、尤 (particularly)、超 (super)
1.2	不甚 (fully)、不勝 (extremely)、好 (fine)、好不 (no better)、颇 (considerably)、颇爲 (quite)、大 (big)、大爲 (much)
1.1	稍稍 (slightly)、稍微 (a little)、稍许 (slightly)、略 (slightly)、略爲 (slightly)、多少 (how much)
0.9	較 (relatively)、比較 (comparatively)、較爲 (more)、還 (also)
0.8	有點 (a little)、有些 (some)

(a) If the number of positive conceptual words is greater than that of negative conceptual words, and the sentiment score of the whole text is more than or equal to 0.6, the text is marked as positive;

(b) If the number of negative conceptual words is greater than that of positive conceptual words, and the sentiment score of the whole text is less than or equal to 0.6, then the text is marked as negative;

(c) If neither of the above rules is applicable, the rule-based classifier marks the text as unknown and uses the low-confidence prediction value of SVM as the final output of the system.

4. Experiments and Results

When people review on tangible products provided by manufacturing industry and invisible services provided by service industry, they often use different vocabulary from different perspectives. Therefore, in order to verify the validity of the improved model comprehensively, this paper chooses product reviews and service reviews in Weibo as the research objects. Firstly, the crawler technology is used to capture 100,000 data information of product reviews represented by apparel and tablet computers from Sina Weibo (www.weibo.com) as experimental data sets. The first 40,000 reviews are selected as training data, and the remaining 60,000 reviews are used as test data. The data sets are analyzed by language rules and enhanced supervised learning. Python language is used to design the language program according to the sentiment analysis system flow designed above. Table 6 shows the results of the analysis and Table 7 compares the results of the two methods.

Table 6: Data List of Analysis Results of Product Reviews.

key = pos_number	value =54209
key = neg_number	value = 5791
key = number_ratio	value = 7.5
key = pos_mean	value = 4.1
key = neg_mean	value = -1.7
key = total_mean	value = 2.8
key = mean_ratio	value = 2.4
key = pos_variance	value = 24843.6
key = neg_variance	value = 1.6
key = total_variance	value = 18047.8
key = var_ratio	value = 15445.8
key = text_pos_number , value = The number of positive microblog reviews is 54209, accounting for 90.3% of all microblog texts.	
key = text_neg_number , value = The number of negative microblog reviews is 5791, accounting for 9.7% of all microblog texts.	

Table 7: Comparison Table of Processing Result of Product Reviews.

Method	positive			negative			mean value		
	P	R	F	P	R	F	P	R	F
N-meta model	0.90	0.31	0.43	0.66	0.70	0.68	0.78	0.51	0.56
An Improved N-Meta Model based on language rules and enhanced Supervised Learning	0.99	1	0.99	0.75	0.95	0.84	0.98	0.99	0.99

In order to illustrate the accuracy of the results in different types of microblog reviews, the crawler technology is used to capture 100,000 service reviews datasets represented by hotel and movie reviews from Sina Weibo (www.weibo.com) as experimental data sets. The first 40,000 review texts are selected as training data, and the remaining 60,000 review texts are selected as test data, using language rules and enhanced supervised learning to train and analyze data sets. Python language is used to design the language program according to the sentiment analysis system flow designed above. Table 8 shows the results of the analysis and Table 9 compares the results of the two methods.

Tables 6 and 8 are the final processing results of two kinds of review data sets designed by Python language according to the flow shown in Figure 2. They show the number, mean and variance of text in both positive and negative classes. In Table 7 and table 9, P represents precise, R represents recall and F represents test value. In

Table 8: Data List of Analysis Results of Service Reviews.

key = pos_number	value = 18976
key = neg_number	value = 41024
key = number_ratio	value = 0.8
key = pos_mean	value = 444.4
key = neg_mean	value = -2.2
key = total_mean	value = 139.7
key = mean_ratio	value = 206.1
key = pos_variance	value = 3708103046.2
key = neg_variance	value = 3.0
key = total_variance	value = 1172792332.7
key = var_ratio	value = 1235481529.3
key = text_pos_number, value = The number of positive microblog reviews is 18976, accounting for 31.63% of all microblog texts.	
key = text_neg_number, value = The number of negative microblog reviews is 41024, accounting for 68.37% of all microblog texts.	

Table 9: Comparison Table of Processing Result of Service Reviews.

Method	positive			negative			mean value		
	P	R	F	P	R	F	P	R	F
N-meta model	0.66	0.70	0.68	0.75	0.95	0.84	0.71	0.83	0.76
An Improved N-Meta Model based on language rules and enhanced Supervised Learning	0.96	0.98	0.97	0.90	0.88	0.89	0.93	0.93	0.93

order to evaluate the validity of the method in this paper, we consider the average value of F-test of positive and negative as the main evaluation criteria. By comparing the standard N-meta model with the improved N-meta model based on language rules and enhance supervised learning, it can be seen that the F-means of the improved N-meta model based on language rules and enhanced supervised learning increased by 0.43 and 0.17 respectively in the two groups of data analysis. Therefore, this shows that the improved model significantly improves the validity of sentiment analysis results, but also can effectively improve the precise of judgment.

The differences in the results obtained from different microblog review datasets may be related to the dictionary of participle, stop words and sentiment words used in the process of text sentence preprocessing and decision score calculation. Whether the vocabulary in the dictionary is comprehensive or not, and whether the corresponding score formulation is reasonable will directly affect the processing results. The improvement of F average value of hotel and film review data set is much less than that of product review data set It is mainly related to the fuzzier evaluation criteria for service products, the more complex and diverse commentary terms, the broader review angle, and the greater difference between training text and test text.

5. Conclusions and Future Research

This paper proposes a system to improve supervised learning for text sentiment analysis of micro-blog reviews: on the basis of traditional sentiment analysis method, language rules are formulated for specific sentence structures, so as to locate sentiment words more accurately and calculate sentiment scores with sentiment dictionary, and then modify classification labels for text with lower decision scores in order to enhance supervised learning. By comparing the improved model with the traditional model, we can draw the conclusion that the model can be improved by processing the special parts of the sentence, such as expressions, conjunctions and conditional clauses, and the sentiment value can be calculated through the sentiment dictionary. This method can significantly improve the effectiveness of sentiment classification results.

Future research will improve the existing methods in the following aspects: 1) adding popular terms and professional terms to dictionaries, 2) correcting the sentiment scores of words in sentiment dictionaries, 3) formulating corresponding language rules considering more complex sentence structures, and 4) incorporating more feature vectors into the research category. At the same time, we can further study the application of sentiment analysis and sentiment analysis in a more fine-grained way in micro-blog texts in a wider field (see He et al. [26]).

Acknowledgements

This study is supported by National Social Science Fund of China under grant No.20BXW013.

References

[1] Abid, Fazeel, Alam, Muhammad and Yasir, Muhammad (2019). *Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter*, Future Generation Computer System-The International Journal of Escience, Vol.95, 292-308.

[2] Abudalfa, S. I. and Ahmed, M. A. (2019). *Semi-Supervised Target-Dependent Sentiment Classification for Micro-Blogs*, Journal of Computer Science and Technology, Vol.19, e06-e06.

[3] Ai, Y., Chen, Z., Wang, S. and Pang, Y. (2018, July). *Recognizing emotions in chinese text using dictionary and ensemble of classifiers*, In Third International Workshop on Pattern Recognition (Vol. 10828, p. 1082807). International Society for Optics and Photonics.

[4] AlAwadhi, S. and Al-Daihani, S. M. (2019). *Marketing academic library information services using social media*, Library Management, Vol.40, 228-239.

[5] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011). *Sentiment analysis of twitter data*, In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).

[6] Balbi, S., Misuraca, M. and Scepi, G. (2018). *Combining different evaluation systems on social media for measuring user satisfaction*, Information Processing & Management, Vol.54, 674-685.

[7] Balijepalli, S. (2007). Blogvox2: A modular domain independent sentiment analysis system.

[8] Baum, M. A. and Potter, P. B. (2019). *Media, public opinion, and foreign policy in the age of social media*, The Journal of Politics, Vol.81, 747-756.

[9] Bergstr`om, A. and Jervelycke Belfrage, M. (2018). *News in social media: incidental consumption and the role of opinion leaders*, Digital Journalism, Vol.6, 583-598.

[10] Bollen, J., Pepe, A. and Mao, H. (2010). *Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena*. WWW 2010, Raleigh, North Carolina, April, 26-30.

[11] Bollen, J., Mao, H. and Pepe, A. (2011). *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena*, In Fifth International AAAI Conference on Weblogs and Social Media.

[12] Bollen, J., Mao, H. and Zeng, X. (2011). *Twitter mood predicts the stock market*, Journal of Computational Science, Vol.2, 1-8.

[13] Cambria, E, Schuller B, Xia Y. and Havasi, C. (2013). *New avenues in opinion mining and sentiment analysis*, IEEE Intelligent Systems, Vol.28, 15-21.

[14] Chaovalit, P., Zhou, L. (2005). *Movie review mining: A comparison between supervised and unsupervised classification approaches*, In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences, HICSS 2005, pp. 112c-112c. IEEE.

[15] Chaturvedi, I., Cambria, E., Welsch, R. E. and Herrera, F. (2018). *Distinguishing between facts and opinions for sentiment analysis: Survey and challenges*, Information Fusion, Vol.44, 65-77.

[16] Cheng, J., Zhang, X., Li, P., Zhang, S., Ding, Z. and Wang, H. (2016). *Exploring sentiment parsing of microblogging texts for opinion polling on chinese public figures*, Applied Intelligence, Vol.45, 429-442.

[17] Chikersal, P., Poria, S. and Cambria, E. (2015). SeNTU: Sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In: Proceedings of the International Workshop on Semantic Evaluation (SemEval 2015).

[18] Chopade, C. R. (2015). *Text based emotion recognition: A survey*, International Journal of Science and Research, Vol.4, 409-414.

[19] Cureg, M. Q., De La Cruz, J. A. D., Solomon, J. C. A., Saharkhiz, A. T., Balan, A. K. D. and Samonte, M. J. C. (2019). *Sentiment Analysis on Tweets with Punctuations, Emoticons, and Negations*, In Proceedings of the 2019 2nd International Conference on Information Science and Systems(pp. 266-270). ACM.

[20] Ding S., Meng M. and Li X. (2014). *Research on Identity Sentence Recognition for Chinese Weibo*, Journal of Information Science, Vol.33, 175-182.

[21] Ellen, J. (2011). *All about Microtext-A Working Definition and a Survey of Current Microtext Research within Artificial Intelligence and Natural Language Processing*, ICAART (1), 2011, 329-336.

[22] Geng, X., Zhang, Y., Jiao, Y. and Mei, Y. (2019). *A Novel Hybrid Clustering Algorithm for Topic Detection on Chinese Microblogging*, IEEE Transactions on Computational Social Systems, Vol.6, 289-300.

[23] Go, A., Bhayani, R. and Huang, L. (2009). *Twitter sentiment classification using distant supervision.* CS224N Project Report, Stanford.

[24] Hasan, A., Moin, S., Karim, A. and Shamshirband, S. (2018). *Machine learning-based sentiment analysis for twitter accounts*, Mathematical and Computational Applications, Vol.23, 11.

[25] Hasan, M., Rundensteiner, E. and Agu, E. (2019). *Automatic emotion detection in text streams by analyzing Twitter data*, International Journal of Data Science and Analytics, Vol.7, 35-51.

[26] He, Y, Zhu, C, Zhu, T and Guo, Q. (2018). *Research on the Emotional Tendency of Hot Topics in Micro-blogs*, Information Studies: Theory & Application, Vol.41, 1-9.

[27] Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R. and Griffith, S. A. (2015). *Social media and disasters: a functional framework for social media use in disaster planning, response, and research*, Disasters, Vol. 39, 1-22.

[28] Java, A., Song, X., Finin, T. and Tseng, B. (2007). *Why we twitter: understanding microblogging usage and communities*, In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (pp. 56-65). ACM.

[29] Jiang X., Xia, K., Xia, X. and Zu B. (2015). *Speech Emotion Recognition Using Semi-Definite Programming Multiple-Kernel SVM*, Journal of Beijing University of Posts and Telecommunications, Vol.38, 67-71.

[30] Kamal, R., Shah, M. A., Maple, C., Masood, M., Wahid, A. and Mehmood, A. (2019). *Emotion Classification and Crowd Source Sensing: A Lexicon Based Approach*, IEEE Access, Vol.7, 27124-27134.

[31] Kaur, H., Pannu, H. S. and Malhi, A. K. (2019). *Multimedia blog volume prediction using adaptive neuro fuzzy inference system and evolutionary algorithms*, Multimedia Tools and Applications, 1-35.

[32] Li, R., Lin, Z., Fu, P., Wang, W. and Shi, G. (2019). *EmoMix: Building an Emotion Lexicon for Compound Emotion Analysis*, In International Conference on Computational Science (pp. 353-368). Springer, Cham.

[33] Lin, D., Li, L., Cao, D., Lv, Y. and Ke, X. (2018). *Multi-modality weakly labeled sentiment learning based on Explicit Emotion Signal for Chinese microblog*, Neuro computing, Vol.272, 258-269.

[34] Liu B. (2012). *Sentiment analysis and opinion mining*, Synthesis Lectures on Human Language Technologies, Vol.5, 158-167.

[35] Liu Z. and Liu L. (2012). *Empirical Classification of Chinese Weibo Based on Machine Learning*, Computer Engineering and Application, Vol.48, 1-4.

[36] López, D. T. (2019). *The Society of the Digital Swarm: Microblogging and Construction of Subjectivity in Homo Digitalis*, In Handbook of Research on Industrial Advancement in Scientific Knowledge (pp. 95-110). IGI Global.

[37] Medhat, W., Hassan, A. and Korashy, H. (2014). *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal, Vol.5, 1093-1113.

[38] Mostafa, M. M. (2019). *Clustering halal food consumers: A Twitter sentiment analysis*, International Journal of Market Research, Vol.61, 320-337.

[39] Ma, L., Liu, X. and Gong, Y. (2016). *Semantic-based Tendency Analysis of Short Texts in Microblog*, Computer Applied Research, Vol.33, 2914-2918.

[40] Neubaum, G. and Krämer, N. C. (2017). *Opinion climates in social media: Blending mass and interpersonal communication*, Human Communication Research, Vol.43, 464-476.

[41] Pan, J., Mou, N. and Liu, W. (2019). *Emotion Analysis of Tourists Based on Domain Ontology*, In Proceedings of the 2019 International Conference on Data Mining and Machine Learning(pp. 146-150). ACM.

[42] Pang, B. and Lee, L. (2008). *Opinion mining and sentiment analysis*, Foundations and Trends in Information Retrieval, Vol.2, 1-13.

[43] Rahimi, Z., Noferesti, S. and Shansfard, M. (2019). *Applying data mining and machine learning techniques for sentiment shifter identification*, Language resources and evaluation, Vol.53, 279-302.

[44] Ramanathan, V. and Meyyappan, T. (2019). *Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism*, In 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC) (pp. 1-5). IEEE.

[45] Rout, J, Choo, K., Dash, A., Bakshi, S., Jena, S. and Williams, K. (2018). *A model for sentiment and emotion analysis of unstructured social media text*, Electronic Commerce Research, Vol.18, 181-199.

[46] Sahni, D. and Aggarwal, G. (2015). *Recognizing Emotions and Sentiments in Text: A Survey*, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.5, 202-205.

[47] Sailunaz, K. and Alhajj, R. (2019). *Emotion and Sentiment Analysis from Twitter Text*, Journal of Computational Science.

[48] Shi, W., Wang, H. and He, S. (2013). *Sentiment analysis of Chinese microblogging based on sentiment ontology: A case study of '7.23 Wenzhou Train Collision'*, Connection Science, Vol.25, 161-178.

[49] Shi, W., Wang, H. and He, S. (2012). *Study on construction of fuzzy Emotion ontology Based on HowNet*, Journal of the China Society for Scientific and Technical Information, Vol.31, 595-602.

[50] Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T. M. (2018). *Big Data: Deep Learning for financial sentiment analysis*, Journal of Big Data, Vol.5, 3.

[51] Tang, D., Zhang, Z., He, Y., Lin, C. and Zhou, D. (2019). *Hidden topic-emotion transition model for multi-level social emotion detection*, Knowledge-Based Systems, Vol.164, 426-435.

[52] Tripathy, A., Agrawal, A. and Rath S. (2016). *Classification of sentiment reviews using n-gram machine learning approach*, Expert Systems with Applications, Vol.57, 117-126.

[53] Turney, P. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*, In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417-424. Association for Computational Linguistics.

[54] Tyagi, Priyanka and Chakraborty, Sudeshna and Tripathi, R. C. and Choudhury, Tanupriya, Literature Review of Sentiment Analysis Techniques for Microblogging Site (March 15, 2019). Available at SSRN, https://ssrn.com/abstract=3403968 or http://dx.doi.org/10.2139/ssrn.3403968

[55] Vo, P. N. and Ngoc, T. V. T. (2019). *Data Mining for Social Network Analysis Using a CLIQUE Algorithm*, In Cognitive Social Mining Applications in Data Analytics and Forensics (pp. 160-187). IGI Global.

[56] Wang, C. H. and Han, D. (2018). *Sentiment Analysis of Micro-blog Integrated on Explicit Semantic Analysis Method*. Wireless Personal Communications, Vol.102, 1095-1105.

[57] Wang, N., Ke, S., Chen, Y., Yan, T. and Lim, A. (2019). *Textual Sentiment of Chinese Microblog Toward the Stock Market*, International Journal of Information Technology & Decision Making (IJITDM), Vol.18, 649-671.

[58] Wang, W. Y., Pauleen, D. J. and Zhang, T. (2016). *How social media applications affect B2B communication and improve business performance in SMEs*, Industrial Marketing Management, Vol.54, 4-14.

[59] Wiegand, M., Balahur, A., Roth, B., Klakow, D. and Montoyo, A.(2010). *A survey on the role of nega-tion in sentiment analysis*, In: Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pp. 60-68. Association for Computational Linguistics.

[60] Wu, X. and Wang, J. (2011). *How about micro-blogging service in China: analysis and mining on sina micro-blog*, In Proceedings of 1st international symposium on From digital footprints to social and community intelligence (pp. 37-42). ACM.

[61] Yao, T. and Lou, D. (2007). *A Study on the Tendency Discrimination of Chinese Emotional Words.* (pp. 222-225). Seventh International Conference on Chinese Information Processing, wuhan:ICCC2007.

[62] Yuan, D., Zhou, Y., Li, R. and Lu, P. (2014). *Sentiment analysis of microblog combining dictionary and rules*, In 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) (pp. 785-789). IEEE.

[63] Yuan, Z. and Purver, M. (2015). *Predicting emotion labels for chinese microblog texts*, In Advances in Social Media Analysis (pp. 129-149). Springer, Cham.

[64] Zhang, S., Wei, Z., Wang, Y. and Liao, T. (2018). *Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary*, Future Generation Computer Systems, Vol.81, 395-403.

[65] Zhang, J., Xia, Y., Ma, B., Yao, J. and Hong, Y. (2011). *Thread cleaning and merging for microblog topic detection*, In Proceedings of 5th International Joint Conference on Natural Language Processing (pp. 589-597).

[66] Zhang, L., Wang, S. and Liu, B. (2018). *Deep learning for sentiment analysis: A survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol.8, e1253.

[67] Zhou, J., Ye, S., Wang, H. (2018). *Application of Convolutional Neural Network in Short Text Emotional Multi-Classification Tagging*, Computer Engineering and Application, Vol.54, 133-138.

[68] Zhou, S., Guan, J., Yu H. and Hu, Y. (2001). *Chinese Documents Categorization Based on N-gram Information*, Journal of Chinese Information Processing, Vol.15, 34-39.

Business School, HuZhou University, Hu Zhou 313000, P.R. China.

E-mail: shiwei@zjhu.edu.cn
Major area(s): Sentiment analysis, text mining, Business intelligence.

College of Business and Public Administration, California State University, San Bernardino, USA.

E-mail: SHe@csusb.edu
Major area(s): Data mining, Electronic Commerce, Business intelligence.

Qiuzhen College, HuZhou University, Hu Zhou 313000, P.R. China.

E-mail: 02246@zjhu.edu.cn
Major area(s): Sentiment analysis, information management, Business intelligence.

(Received May 2020; accepted June 2021)