A Study on the Optimization of Text Emotional Classification Based on Class Representation Words and Naive Bayes

Hongyu Liu
E-mail: liuhongyu777@163.com

College of Economics and Management, Heilongjiang Bayi Agricultural University, Daqing, China

Abstract—There is a certain similarity of the characteristic vectors of the comment statements between the comments between the commentary emotion category, bringing certain difficulties to text emotion 3 classification. This article proposes a specification of the specification. It is mainly from 8 aspects of fresh products to 8 aspects (emotional adjustment words + feature words). Add all kinds of representatives selected to a word-defined dictionary, by changing the word words in the comment statement, in turn increases the difference between the comment statement feature vector between the category during the quantization process, thereby enabling classification Easy to identify the emotional polarity of comment statements. After verification, this method achieved a good classification effect, the correct rate reached 86.14%, the recall rate reached 84.97%, which increased 4.24% and 4.42% compared to the method without dictionary.

Keywords- text mining; naive Bayes; classification

1 Introduction

The development speed and scale of Chinese e-commerce have reached an unprecedented level, and the total transaction in 2020 has reached 29.16 trillion yuan. This promotes the development of commercial society, and also drives supply chain enterprises to perform digital changes. Such as Tmal, Jingdong et al. e-commerce platform have attracted more and more consumers' shopping behavior. Consumers, who are active on the platform, also leave a massive data, such as consumers' ID, order time, product reviews, etc., which contain huge commercial value wait for data mining.

There are two main channels for e-commerce product comments. The first is consumers' spontaneous comment, who express their real idea for products and services, which are valuable for seller and other consumers.

The second way is incentive strategies from the platform, which inspire consumers to participate in adding comments, for that they can get rewards for shopping. Such as Jingdong beans rules, which are operated as follow: you can get 0 Beans with less than 10 words, get 10 beans when the number of reviews is more than 10, if the goods are sold between 20 and 100, and get 20 beans if the price is greater than 100. However, the content of the comment is from the copy, and the copy ratio exceeds more than 80% (subject to the number of words), or the punctuation is too much, and the evaluation content will not give beans. Strategies such as this are used to improve the usefulness of consumer comments.

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2327117

Business and researchers often use text mining method to analyze the consumer comments, especially for textual emotional classification. After emotional classification of comments, we can find factors that make consumers satisfaction. Finally, the comparison of the three type emotional classification can also be performed, and more information with reference value can be performed. According to the processing of the above to the text, reference information can be provided to the corresponding managers. However, the accuracy of the existing method needs to be further improved. This article will join the vocabulary dictionary to the Bayesian classification model to improve classification accuracy.

2 Text sentiment classification method

2.1 Machine learning methods

There are two main methods for text sentiment analysis. One is based on the sentiment dictionary method, which calculates the sentiment score of the text based on sentiment words to determine the sentiment tendency of the text. On the one hand, this method has a large dependence on emotional words, on the other hand, The same emotional word will express different emotions due to the presence or absence of modifiers, and these factors will cause large errors in the emotional judgment of the text. Therefore, in the past few years, the method of emotional dictionaries has been used to classify the sentiment of texts less frequently. Another is a method based on machine learning, which is dedicated to studying how to use computing to improve the performance of the system itself through experience [1]. First by constructing a suitable algorithm model, and then transmitting the empirical data to the model, by learning the inherent laws of the data, the model will provide us with corresponding judgments when facing new similar data. Machine learning methods are divided into two types, supervised and unsupervised. In supervised learning, each training and verification sample is composed of an input object (generally a feature vector) and an expected output value. Compared with the expected output value, the classification effect of the model can be verified. Unsupervised learning refers to that the input data of the model has only feature vectors and unlabeled expected output values, and uses various unsupervised algorithms to identify the internal laws of the data and classify them according to the internal connection.

In recent years, machine learning-based methods have been widely accepted for emotional classification of texts [2]. Calculate their PTF-IDF by selecting nouns, adjectives, etc., and sort them. After setting the threshold, filter out the domain sentiment dictionary, and use this dictionary to assist text sentiment classification. He and Zhang et al. constructed an emotional word network of word-word, word-object relationships, the characteristics of the text are extended correspondingly in the weight and feature set through the emotional word network, and then the machine is used to classify the emotion of the text [3]. Wang et al. applied a high-dimensional mixed feature method was proposed to classify text sentiment into three categories, and the classification accuracy rate reached 0.8469 [4]. Liu et al. analyze and research the characteristics of micro-blog, by utilizing multiple features of micro-blog text under the topic, established the micro-blog sentiment polarity classification model, judge the polarity of micro-blog by adopting a classification of machine learning, uses the relationship between the repostment, the comment and the praise of micro-blog, the number of fans and the number of concerns to implement graph-based optimization, and propose a method of micro-blog's multi feature sentiment polarity classification based on the topic of micro-blog, and they found that this method has a favorable effect on sentiment polarity classification of micro-blog, and the classification effect is 0.8630 [5]. Yan et al. proposed a model of analyzing stock text based on emotional dictionary and LDA, and constructs a comprehensive emotional dictionary through stock proprietary words, degree words, and transfusion words, and then use LDA to calculate documents for emotional words - theme, document - probability distribution, calculate the correlation between the comment statements, etc., thus judge the sentence emotion [6]. Qin Feng et al achieves text feature information extension by establishing a text topic, then on the topic and topic, the text feature information extension is achieved, this method has achieved good results in text emotion 3 category [7].

This article uses machine learning method for supervision. We first make emotional note on a large number of comments, and then establish a model classifier to learn about the text of the annotation. During this time, we reach the purpose of classifier optimization by optimizing text characteristics, adjusting parameters, thereby save the classifier when the classification achieve a purposed effect. Finally, we call the classifier to emotionally classify similar text.

2.2 Naive Bayes Classifier

Naive Bayesian classifier is one of the commonly used methods in machine learning. It is a supervised learning algorithm. In recent years, it has often been used for text sentiment classification. This article mainly uses Naive Bayes algorithm to review e-commerce shopping. Perform emotion 3 classification.

It calculates the category of text x based on class prior probability and attribute conditional probability, and its expression is shown in (1):

$$H(X) = {}^{\arg\max}_{c\in y}P(c_j)\prod_{i=1}^{d}P(x_i|c_j) \tag{1}$$

Where $p(c_j)$ is the prior probability of category c_j, $P(x_i|c_j)$ is the sub-key probability of attribute x_i in category c_j.

A prior probability $P(C_j)$ refers to the ratio of a total number of categories accounts for all categories. Suppose 3 categories in the training concentration D, a total of M samples, where c_j accounts for N (N < M) sample, as shown in (2):

$$P(c_j) = \frac{N}{M} \tag{2}$$

The post-test probability refers to the probability of a feature in a certain dimension of the sample feature vector, assuming that the category c_j has Q sample, and the number of q times the characteristic x_i appears at the nth dimension of the feature vector, which is shown in (3):

$$P\left(x_i\middle|c_j\right) = \frac{q}{Q} \tag{3}$$

Suppose a feature of the verification set and the concentration of the to be tested have never appeared in the training set q = 0, then $P\left(x_i\middle|c_j\right) = 0$, will cause H(X) to zero, so other dimensional characteristics tend to category c_j, and the result will return him as other categories. In order to avoid the occurrence of such phenomena, “smoothing” processing is usually necessary to estimate the probability value. “Laplas Smooth” and “Lidstone Smooth” are the most common way of use. When a=1 causes “Laplas Smoothing”, when 0 <a <1, we use

“LidStone Smoothing”, where a represents the number of possible categories in the training set D, B is characterized by the characteristic of the category $\mathrm{c_j}$. The number of types of $\mathrm{x_i}$, as shown in (4) and (5).

$$\mathrm{P}' = \frac{\mathrm{N+a}}{\mathrm{M+A}} \tag{4}$$

$$\mathrm{P}'\left(\mathrm{x_i}\middle|\mathrm{c_j}\right) = \frac{\mathrm{P+a}}{\mathrm{Q+B}} \tag{5}$$

3 Experimental process and analysis of experimental results

3.1 Comment data collection and processing

This study uses python packet capture tool to collect more than 12,000 Cherries comments from Jingdong fresh Mall, from 2019 to 2021. After deleting duplication process, comments with less than 4 words in the sentence are deleted, and 10362 valid comments are retained. These 10362 comments are tagged with artificial emotion categories The emotional polarity label rules are shown in Table 1. As long as the corresponding words appear in the comment sentence, and it obviously feels like a positive, middle or negative review, this comment will be classified into the corresponding category. The situation of each category after labelling is as follows Table 2.

Table 1. Comment emotion polarity label rules

category	rule
Praise	Very satisfied, satisfied, good, coming back next time, fresh, very fresh ...
Average	Fair, OK, OK, not too fresh, not very satisfied, a little disappointed ...
Bad review	Disappointed, unpalatable, not fresh, rotten, bad, thrown, small and sour ...

Table 2. Comments classification summary

comment category	Comments expected		
	Praise	Average	Bad review
total	5008	1922	3432

The purpose of segmenting the comment text is to make the feature of the comment text more vectorized, which is helpful for the classification of the comment text. This article uses “staple” segmentation. The process of segmentation and vectorization is as follows.

Example 1: “Logistics are fast and cherries are also fresh”.

The participle is as follows: “logistics / soon /, / cherries / also / very / fresh /”.

The corresponding feature vector is “[0 0 0 1 0 1 1 1]”.

Example 2: “I received something, it's still fresh, and the taste is not good” .

The participle is as follows: “something / received / received /, / still / reported / fresh /, / taste / not / good”.

The corresponding feature vector is “[1 1 1 0 1 1 0 0]”.

3.2 Emotion classification of traditional review text

This article uses 80% of the comment text as the training set and the remaining 20% as the validation set, as shown in Table 3:

Table 3. Comment classification training set and validation set

comment category	Comment text			
	Positive	Middle	Negative	total
Training set	4005	1554	2741	8300
Validation set	1003	368	691	2062
total	5008	1922	3432	10362

(1) Evaluation index of classification effect, as in (6) and (7):

$$P=\frac{T_P}{T_P+F_P} \tag{6}$$

$$R=\frac{T_p}{T_P+F_N} \tag{7}$$

Where P is the accuracy rate and R is the recall rate. T_P Number of positive reviews for the classifier, F_P Score the difference to the number of positive reviews for the classifier, F_N score of negative reviews for the classifier.

(2) Parameter setting

The Naive Bayes classifier is used to train and verify the review text. Considering that some features in the validation have never appeared in the training set, the final probability is 0, the α value needs to be adjusted to avoid “Lapras Smoothing” (α = 1) or “LidStone Smooth” (0 <α <1) is used.

(3) Analysis of classification results

By adjustingαValue, the resulting classification effect is shown in Table 4. below:

Table 4. Validation set classification results

α=0.2		α=0.4		α=0.6		α=0.8		α=1	
P	R	P	R	P	R	P	R	P	R
0.8185	0.7895	0.8205	0.7978	0.8175	0.7997	0.8190	0.8055	0.8161	0.8065

It can be seen from the p and r values in Table 4 that the classification effect is not very good. In order to see the classifier for each category for more detail, it can be visualized using the formula of the recall rate, and we show it in α= 1, as shown in Table 5.

Table 5. Recall rate of positive, middle and negative reviews

Flag emotion categories	Recall
Positive	0.8395
Middle	0.7120
Negative	0.8090

It can be seen from Table 5 that the recall rate of the middle review is only 71.20%, which means that 28.80% of the middle review is identified by the classifier as other categories, and 19.1% of the negative reviews are classified by the classifier as other categories. 16.05% of the positive reviews were classified into other categories, indicating that the classification effect is not good and needs to be improved.

3.3 Classification of lexical text sentiment classification

(1) Analysis of text features between emotion categories

It can be found from Table 4. and Table 5. in section 3.2 that the classification effect is not very good. After observation and research, it is found that one of the reasons for the above phenomenon is that some sentences in the positive, middle, and negative evaluations have been segmented There is a high degree of repetition of words between category sentences, causing their feature vectors to have a high similarity, which eventually causes the classifier to misclassify. For example:

Example 3: “Logistics is fast and cherries are also fresh”.

Example 4: “Logistics is not very fast, cherries are a little fresh”.

The comment sentence that are segmented are shown:

Example 3: “logistics /very fast /, / Chelizi / too / very / fresh /”.

Example 4: “logistics / not /very fast /, / Chelizi / a little / not / fresh /”.

The repeated words in Example 3 and Example 4 after word segmentation are “logistics”, “fast”, “Chelizi”, and “fresh”.

The result of the summary quantization is:

Example 3: “[0 1 1 0 1 1]”

Example 4: “[1 1 1 1 1 1]”

Due to the high degree of repetition of words between Example 3 (good) and Example 4 (middle review), their feature vectors are highly similar, which leads to misclassification during classification.

The reason for the misclassification is that some emotion modifiers and feature words are separated during word segmentation. For example, the emotion modifiers for logistics speed in Example 3 and Example 4 are "very" and "not very" respectively. In Chinese, there are obvious differences between them. However, after the word segmentation, the emotional adjustment word “very” and the characteristic word “fast” in Example 3 are grouped together, while in Example 4 only the partial adjustment word “very” and the characteristic word “quick” is grouped together, and the remaining emotional modifiers “not” are separate words, which promotes the repeated text “fast” in the two sentences. Such word segmentation has a very high feature vector after the text vector. Similarity, the semantics of the machine is weakened, adding difficulty to text classification.

In order to change the similarity between the review texts caused by improper emotion modifiers and feature word segmentation, this paper selects the class representative words (emotional modifiers and feature words, hereinafter referred to as: class representative words) of various reviews as the custom dictionary of jieba word segmentation increases the difference between text feature vectors by changing the word segmentation mode of emotion adjustment words and feature words. The specific process is as follows:

Example 5: “Logistics is very fast, and cherries are also very fresh”.

Example 6: “Logistics is not very fast, and the cherries are a little not fresh”.

After adding a custom dictionary (the words in the dictionary are: logistics are fast, very fresh, not fast, and a little new), the word segmentation is performed, and the word segmentation is as follows:

Example 5: “Logistics/ very fast /, / cherries / also / very fresh /”.

Example 6: “Logistics / not very fast /, / cherries / not fresh /”.

The result of the summary quantization is:

Example 5: “[0 1 0 0 1 1]”

Example 6: “[1 0 1 1 0 1]”

From the above feature vector, we can see that after using class representative words as a custom dictionary for word segmentation, the text feature vectors of Example 5 and Example 6 will have a large difference. The example 5 will be more easily classified, and the classifier is recognized as a positive review, and the example sentence 6 belonging to the middle review will be more easily recognized as a middle review by the classifier.

Through the above experiments, it can be shown that in text sentiment classification, the representative words of each class should be found to the largest extent, and use it as a custom dictionary of word segmentation, so as to increase the differences between various types of text feature vectors. It is easier to classify into the category it should belong to.

(2) Establishment of initial class representative word dictionary

This experiment mainly looks for class representative words from the following 8 aspects, which are: quality, taste, size, logistics, consumer intentions (clearly indicating whether they will buy in the future), overall evaluation, customer service, others. For example, in terms of quality, “a little fresh” and “very fresh” are representative words of praise, “not fresh”, “mildewed”, and “rotten” can be representative words of negative reviews, but like “not fresh enough”, “freshness in general”, etc. can be the representative words of the middle review. According to such a rule, find other representative words, and add these representative words to the complete dictionary u.

In the manual labelling of emotional sentiment of text, there is no strict boundary, and the representative words in one category may appear in a small number of other categories, especially between the middle and positive reviews, and between the middle and negative reviews. So even if you find class representative words, some of which do not have any promoting effect on classification, on the contrary, they may also play a negative role. We need to delete the negative class representative words. In order to solve this problem, this article uses a subset dictionary method, that is, outputting a subset dictionary u from the complete set of $\mathrm{A_i}$. Considering that there are too many class representative words in the complete set U, the number of output subset dictionaries is huge (assuming 100 words in U, then the number of the subset is 2^{100}), which brings huge difficulties to the data calculation of the experiment. Therefore, the number of subset words taken out each time in this experiment is n-1 (the number of words in the complete set u is n). In this way, the number of subsets output each time is only n, the data operation time is significantly reduced. After the operation, each subset dictionary will output a corresponding classification result. We take the best from subset U and return the dictionary corresponding to it. We use this dictionary as the complete set dictionary U, then output its subset dictionary, use the subset dictionary as the segmentation custom subscript, and then calculate the corresponding classification results after segmentation. In this process, the experiment goes on and on until the result stops rising. This indicates that words with negative effects on classification have been removed from the complete dictionary U. This sub-dictionary can be used as the final custom dictionary for stuttering word segmentation. The initial complete set of classes selected by this experiment is shown in Table 6.:

Table 6. Class representative words about product features

product features	Class representative words
quality	Not fresh, not too fresh, rotten, rotten, not fresh enough, moldy, not very fresh, broken, all broken, broken, rotten, somewhat broken, a few bad fruits, not very fresh, quite Fresh, fairly fresh, a little stale, not too fresh, average freshness, a little sloppy, very fresh, very fresh, no bad, no one bad, no bad
Taste	No flavor, bad taste, too unpalatable, unpalatable, unpalatable, not so sweet, average taste, average taste, a little sour, light, not very sweet, not too sweet, not too tasty, one A little sweet, not very sweet, not very tasty, good taste, sweet, especially sweet
Head	Small and sour, relatively small, a little small, average size, really big, very large, big, big, sweet, big
Logistics	Too slow, very slow, logistics too slow, wait a few days to arrive, 4 days to arrive, no delivery, fast, fast, not fast, logistics is not fast, a little slow, fast delivery, fast Fast cargo, fast logistics, fast logistics, fast logistics, express delivery, logistics delivery, fast, very fast
Future trip For intention	No more, no more, no more, recommended buy, still buy, come back next time, continue again, come back, will come again, buy again next time
overall evaluation	Uncomfortable, unreliable, okay, okay, okay, okay, too general, not too strong, still okay, overall okay, not very satisfied, very good, too liked, not bad, Like it very much, very satisfied, very satisfied
Customer service	No reply, good attitude, good attitude, good service, good service, conscientious boss, conscientious merchant, good merchant, responsible
other	Throw away, not good, lost, not eaten, thrown, too much, cannot eat, a little more expensive, still not enough, a little, a little, a little, not too much, very practical

(3) Optimization of class representative dictionary

The class representative words selected in this experiment are shown in Table 6., including 125 words. The result of the first operational output is 124, the corresponding 124 classification results are also calculated. The best classification accuracy among the classification results is 0.8563, and the corresponding subset dictionary is output A_i.

We use this dictionary A_i as a complete set of dictionary U, compute its subset dictionary A_i.

After the reciprocating word, quantization, emotional classification, the continuous output subset dictionaries are generated, which can be used as a custom dictionary. Until the classification result does not rise, as shown in Fig. 1. In this experiment, when the number of subset words is reduced to 118, the classification result is not changing, indicating that there is no interference classification in the dictionary. The calculated subset dictionary is used as a custom dictionary of a word.

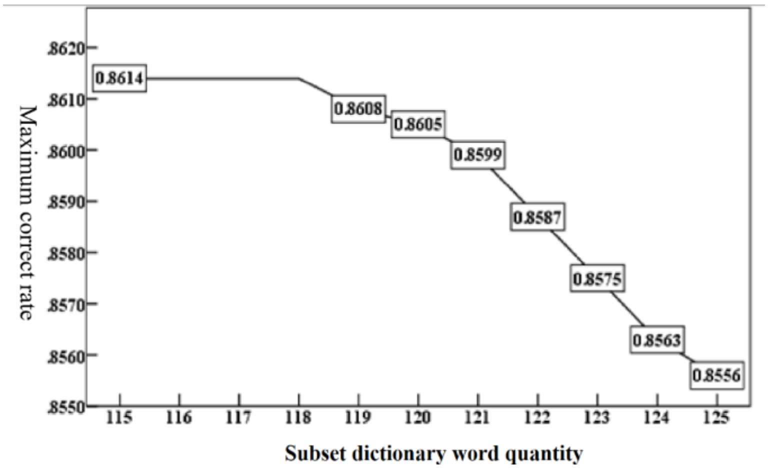

Fig.1 Subset dictionary classification correct rate

(4) Text emotion classification with class representative dictionary

We selected the number of subset dictionary words from 115 to 118 as shown in Fig. 1, and the maximum correct rate of all the subset dictionaries is 0.8614. We use this dictionary as a custom dictionary for stuttering word segmentation to classify comment texts. The classification result are shown in Table 7.

Table 7. Classification of lexical text

α=0.2		α=0.4		α=0.6		α=0.8		α=1	
P	R	P	R	P	R	P	R	P	R
0.8493	0.8220	0.8546	0.8332	0.8579	0.8414	0.8614	0.8497	0.8545	0.8468

(5) Comparison of experimental results with and without class representative words

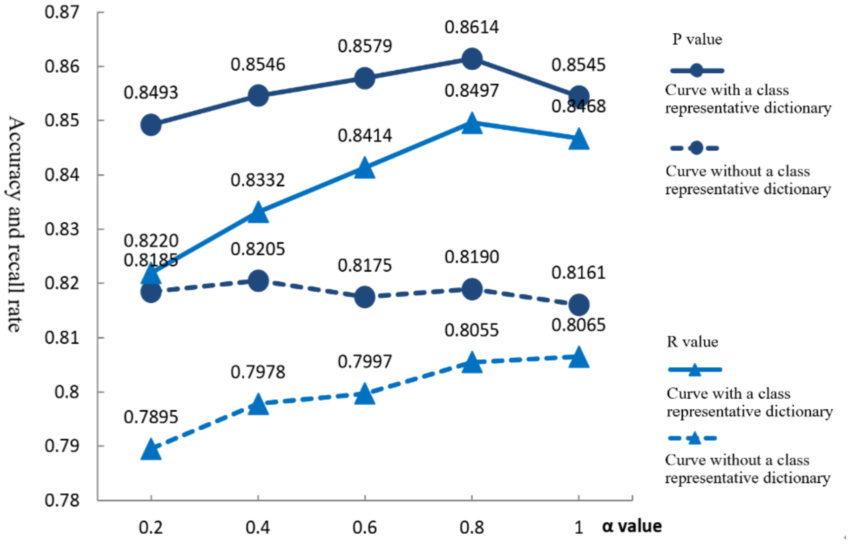

Fig. 2 P and R value change comparison diagram in adding the class dictionary

We can see from Fig. 2 that after adding the class representative words, the accuracy rate and recall rate have improved significantly. The best accuracy rate is 86.14%, a relative increase of 4.24%, a recall rate of 84.97%, and a relative improvement of 4.42. %. The results show that the class representative morphology used in this paper can obviously promote the text 3 type classification experiment.

4 Conclusions and Outlooks

4.1 Conclusions

This paper mainly studies the emotional classification of e-commerce comment texts. The traditional direct classification needs to be further improved, especially for the emotion classification of three polarities. Based on this, this paper proposes a class representation morphology, which reduces the recognition difficulty of the classifier and increases the accuracy of emotion classification by increasing the difference of feature vectors between comment categories.

4.2 Outlooks

In this study, class representative words are found manually. This process takes a lot of time, which affects the efficiency of classification calculation. In the next research, we will automatically select the representative category words from the maximum range through the comparison between category comment texts and machine learning method, so as to classify the emotion of comment sentences. The next research content will help to improve the classification efficiency, and the classification effect will be better.

Acknowledgments. This work was supported by Heilongjiang Bayi Agricultural university science training project under grant (RRCPY201804), and supported by Heilongjiang Bayi Agricultural university PhD Talent Scientific Research Support Project under grant (XDB201810).

References

[1] Zhihua Zhou, Machine learning. Beijing: Tsinghua University Press, 2016, pp.188-193.

[2] Xiao Wu, Lei Wang, “Investigation on Sentiment of Reviews with Shopping Field Dictionary Construction,” Computer Technology and Development. vol. 27, pp. 194-199, July 2017.

[3] Tianxiang He, Hui Zhang, Bo Li, Chunming Yang, Xujian Zhao, “Sentiment classification combined with sentiment lexicon network for Chinese short text,” Application Research of Computers. vol. 32, pp. 2905-2909, Oct. 2015.

[4] Yizhen Wang, Xiao Zheng, Dun Hou, Hao Hu, “Short Text Sentiment Classification of High Dimensional Hybrid Feature Based on SVM,” Computer Technology and Development. vol. 28, pp. 88-93, Feb. 2018.

[5] Zhiyuan Liu, Junbo Gao, “Multi-feature based sentiment orientation identification for Micro-blog topics,” Information Technology and Network Security. vol. 36, pp. 60-62, Aug. 2017.

[6] Feng Yan, Tengfei Du, “Sentiment analysis of stock market text based on sentiment dictionary and lda model,” Electronic Measurement Technology. vol. 40, pp. 82-87, Dec. 2017.

[7] Feng Qin, Chao Huang, Xiao Zheng, Guangmei Shao. “Topic-based Contextual LSTM for Short-text Sentiment Classification,” Journal of Anhui University of Technology (Natural Science). vol. 34, pp. 289-295, July 2017.