Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030304
ISSN 2790-0983 Vol. 3, Issue 3: 27-34

A Hybrid Semantic Understanding and Machine Learning Based Algorithmic Framework for Chinese Text Sentiment Classification

Xuehang Shao*

Heilongjiang University of Industry and Business, Harbin 150025, China

46176838@qq.com

**corresponding author*

Keywords: Semantic Understanding, Machine Learning, Chinese Text Sentiment, Sentiment Classification Algorithm

Abstract: Text sentiment classification is mainly used to determine the positive and negative aspects of sentiment, and obtains practical applications in user selection, information query, and information screening. In order to address the shortcomings of existing Chinese text sentiment classification algorithms, this paper briefly discusses the processing and labeling of the corpus and data collection for the implementation of the classification algorithm framework, based on the discussion of the steps of Chinese text sentiment classification based on semantic understanding and the process of Chinese text sentiment classification based on machine learning. In addition, the design of a hybrid semantic understanding and machine learning based Chinese text sentiment classification algorithm framework is discussed, and the experimental tests on sentiment classification of Chinese text by this paper's algorithm and DF, IG, and SVM are conducted, and the experimental data show that the check-all rate, check-accuracy rate, and F-measurement rate of this paper's algorithm are as high as 87.12% on average. Therefore, it is verified that the hybrid algorithm based on semantic understanding and machine learning is of high practical value in Chinese text sentiment classification.

1. Introduction:

In the face of different text data, there are still many problems in the practical application of how to discriminate the sentiment tendency of text and the application area of sentiment has been diversified. Therefore, a more efficient and scientific approach is needed to solve the problem.

Nowadays, more and more scholars have done a lot of research in Chinese text sentiment

classification algorithm through various techniques and system tools, and some research results have been achieved through practical research. Tubishat proposed a Chinese text sentiment classification method combining machine learning model and semantic understanding. This method extracts different sentiment features from Chinese text datasets and uses them as input to a classification model combining machine learning model and semantic understanding. Based on the machine learning model and semantic understanding technology, the classification results of this method and the classification results of SVM are experimentally analyzed, and the classification effect of this method is optimal [1]. Lee proposed a data balancing scheme for text sentiment classification. The core algorithm of the scheme can remove some majority texts near minority texts to balance the class distribution of data in locally dense mixed regions. The machine learning algorithm SVM is applied to 8 imbalanced Chinese datasets, and the effectiveness of the proposed method is verified. The experimental results show that the LDMRC+SS and LDMRC+RS methods outperform the corresponding LDMRC methods on the Chinese dataset. This shows that simply using local boundary cutting cannot achieve the best effect, and text sentiment classification needs data rebalancing strategy [2]. Tubishat, with the help of word2 vec, uses neural networks in Deep Learning (CNN) and long and short-term memory artificial neural networks (LSTM) to classify sentiment on balanced datasets. The improved model is obtained by comparing with the classification method. An improved balanced cross-entropy loss function for text sentiment classification is proposed for balanced datasets. Experiments show that this model has the highest classification accuracy and the shortest training time. In the case of fewer positive samples, its recall rate is as high as 88.9%, and the classification performance of the modified classification model is better than that of the traditional model [3]. Although the existing research on Chinese text sentiment classification algorithm is very rich, it still has some limitations in real practice.

In this paper, we start from the concept of text sentiment classification, analyze four classification steps in semantic understanding Chinese text sentiment classification, summarize three classification models for machine learning Chinese text sentiment classification, namely SVM classification method and plain Bayesian classification method, and introduce the feature weight calculation process of Chinese text sentiment classification, and propose a classification algorithm framework combining semantic understanding and machine learning, which The method overcomes the difficulty of manual annotation of the corpus based on machine learning, and improves the classification accuracy by integrating the classification of corpus samples according to the classification tendency of both methods. The experiments prove that the method does have the desired effect.

2. Hybrid Chinese Text Sentiment Classification Algorithm Based on Semantic Understanding and Machine Learning

2.1. Semantic Understanding Chinese Text Sentiment Classification

The approach based on semantic understanding is roughly divided into the following steps:

(1) Constructing an initial sentiment word lexicon. The initial dictionary of sentiment words that can be used as sentiment words in Chinese text sentiment classification training is obtained by manually screening the sentiment words in the collected relevant text database [4].

(2) Determination of emotion verb weights. Due to the different text information, there will be some differences in the positive and negative phase paper of its expression, and the text word weights are obtained in different ways [5].

(3) Text preprocessing. The data samples to be classified are divided into words and lexical annotations, after which the sentiment words and phrases suitable for training of the hybrid classification model are selected according to the corresponding principles [6].

(4) Text sentiment classification. The calculation of the weights of sentiment words and phrases can calculate the thought value and sentiment weight of the work, and thus determine the positive and negative nature of the data [7].

2.2. Machine Learning Chinese Text Sentiment Classification

Based on the machine learning emotion classification problem, its classification process mainly includes the following aspects.

(1) Text preprocessing

Text preprocessing is the process of cleaning and preparing data for text classification, which is a necessary stage for converting semi-structured or unstructured text into an appropriate text representation [8]. The usual operations for Chinese go through the following steps: word separation, deactivation, etc. [9].

(2) Feature selection

The main idea is to extract various feature values from the text to be processed and combine them into a set that represents the information carried by the text [10].

(3) Feature weight calculation

In a text, TFIDF combines the word frequency (TF) and the inverse document frequency (IDF), and the feature weights are calculated by the following formula:

$$Q_{vx} = fk_{vx} * \log(\frac{m}{ck_x}) \tag{1}$$

In the above equation, fk_{vx} denotes the number of occurrences of the x th sentiment item in the v th text, m denotes the total number of texts in the training set, ck_x denotes the total number of texts containing the x th sentiment item, and Q_{vx} denotes the TF-IDF weight corresponding to the x th sentiment item in the v th corpus sample [11].

(4) Classification models

1) Support vector machine (SVM): its core idea is to feed the nonlinearity of the low-dimensional input space data to the high-dimensional number of approximate subtraction kernel space, and find the partition space that meets the classification conditions in the low-dimensional number of approximate subtraction kernel space, so that the classification space of its two-sided training points is maximized [12].

2) Plain Bayesian classification method

In the case of Chinese text sentiment classification, the plain Bayesian classification method is shown in Equation (2) [13].

$$k_{DH} = \arg\max G(k_y) \prod_x G(b_x | k_y) \tag{2}$$

Where $G(k_y)$ is the prior probability of category k_y and $G(b_x | k_y)$ is the posterior probability of feature b_x in category k_y. For Chinese text sentiment classification, the specific process of the algorithm needs to be improved [14]. Defining that the category of text $c = \{q_1, q_2, ..., q_m\}$ belongs to $A = \{a_G, a_D\}$, and considering the weights of sentiment words under the condition that sentiment features do not affect each other, the process of implementing this classification algorithm is shown in Equation (3) [15].

$$a_{DH} = \arg\max\left\{G(a_y)\prod_{x=1}^{m} G(q_x, a_y)^{qk(q_x)}\right\} \tag{3}$$

$G(a_y)$ is the prior probability of sentiment classification in category a_y, $G(q_x, a_y)$ is the probability of sentiment classification after feature word q_x in category a_y, $qk(qx)$ is the weight of word q_x in the test corpus, and $qk(qx)=1$ when plain Bayesian subtype weights are used.

3. A Survey Study on the Framework of Chinese Text Sentiment Classification Algorithm Based on a Mixture of Semantic Understanding and Machine Learning

3.1. Data Collection

Data collection experiments are usually conducted on Chinese datasets. The Chinese data collection uses a university corpus of text categories, and five kinds of corpus are selected from them, including 245 in finance and economics, 234 in sports, 128 in science and education, 321 in law, and 265 in Internet information, which is an unbalanced data collection [16]. In the test, three-fifths of the texts in each category were randomly selected as the training set, and the rest of the texts were used as the test set, and the test was cycled four times, and the average value was taken as the experimental result [17].

3.2. Processing and Annotation of the Corpus

(1) Chinese word separation processing of the corpus

In this paper, a maximum matching algorithm is used to classify the Chinese text in the corpus, and a sentiment dictionary is used as the basis for feature selection [18].

(2) Emotion annotation of the corpus

Texts with a composite score greater than or equal to 5 are labeled as positive texts, and those with a score less than 6 are labeled as negative texts, which are used as the training corpus. The detailed data of the corpus are shown in Table 1.

Table 1. Test corpus data

Corpus	Quantity	Positive	Negative	Chinese feature words	Chinese emotional word
1	4239	874	615	1987	763
2	4612	898	1087	1653	974
3	4672	1123	941	1793	815
4	4313	1265	810	1523	715

From the three data corpora, 418 data samples with an average score of 3 and 1254 data samples with a score of 6 were pooled and selected as the training corpus, and the training samples were composed in proportion to the data samples of affective tendency features, and their detailed data are shown in Table 2.

Table 2. Training corpus data

Corpus	Quantity	Positive	Negative	Chinese feature words	Chinese emotional word
1	1514	321	418	286	489
2	2549	654	418	976	501
3	2921	1254	418	873	376

4. A Hybrid Semantic Understanding and Machine Learning Based Chinese Text Sentiment Classification Algorithm Framework Application Study

4.1. A Hybrid Semantic Understanding and Machine Learning Based Chinese Text Sentiment Classification Algorithm Framework

In this paper, a self-supervised classification model based on a combination of semantic understanding and machine learning is used. The whole classification process is divided into two parts, and the specific process framework of the algorithm is shown in Figure 1.

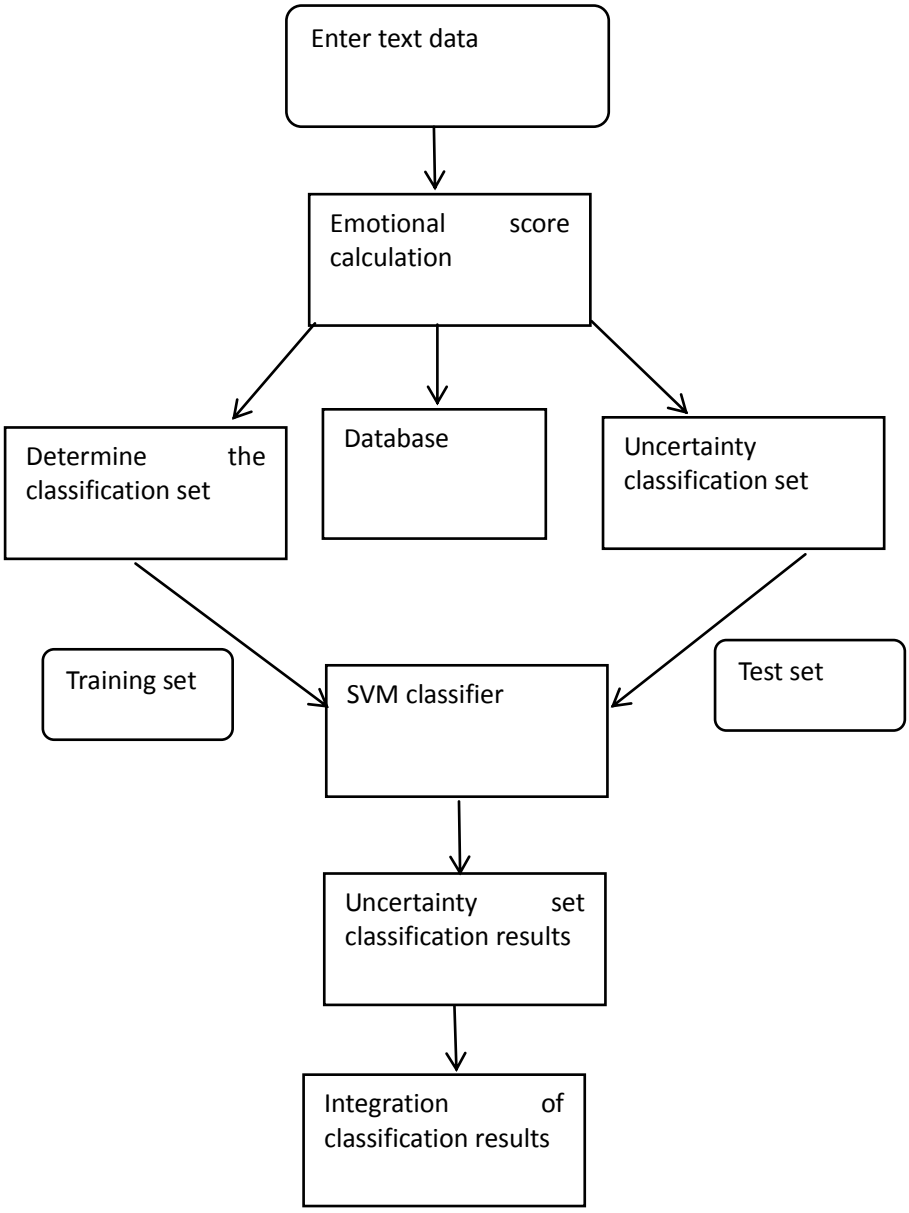

Figure 1. Hybrid Chinese text sentiment classification model based on semantic understanding and machine learning

The hybrid Chinese text sentiment classification based on semantic understanding and machine learning mainly includes the following aspects.

(1) In the first stage, the text is divided into words and deactivated, and then the score of each

sentiment word is obtained by using the baseline words and the database based semantic similarity calculation of words. The rest of the texts are put into the uncertain classification set.

(2) In the second stage, we build a machine learning module based on the classification results of the first stage. Firstly, we use the sentiment dictionary for feature selection, and then use the probabilistic semantic potential model to downscale and semantic association process the classifier, and use the data in the definite classification set as the training set, and reclassify the uncertain classification set.

4.2. Application of a Hybrid Semantic Understanding and Machine Learning Based Chinese Text Sentiment Classification Algorithm Framework

In this experiment, 600 43 texts in the training sample set were classified by semantic understanding, and then the first 300 texts with obvious sentiment tendency were selected as the training texts for SVM, and then four sets of data in the test corpus were used to verify the performance of the classifier. In the feature selection process, DF, IG, SVM and this paper methods were used to classify Chinese text sentiment, and the experimental results obtained from each group were compared separately. The detailed results are shown in Table 3:

Table 3. Algorithm performance analysis data

Sample	Search completion rate	Check accuracy rate	F-measurement
SVM	81.89%	85.17%	83.97%
IG	83.87%	84.76%	85.12%
DF	83.65%	84.12%	82.89%
Methodology of this article	86.78%	88.89%	87.98%

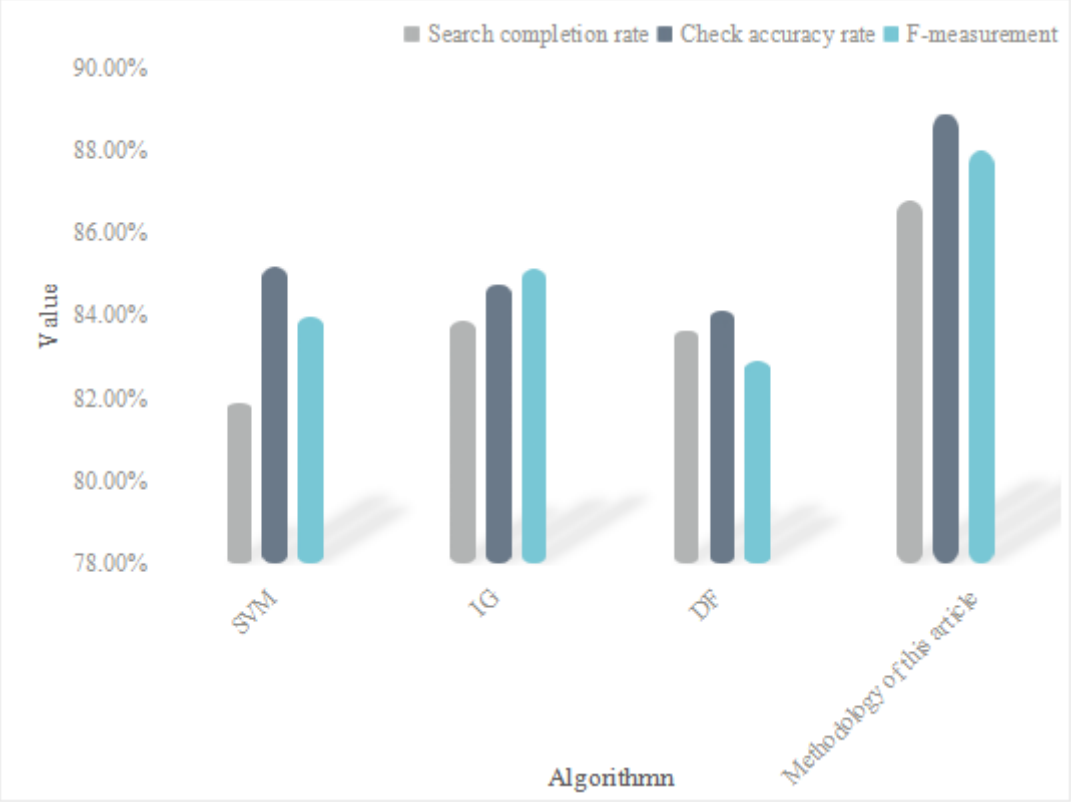

Figure 2. Comparison of classification performance analysis of algorithms

As can be seen from the data in Fig. 2, although a hybrid approach of semantic understanding and machine learning was used to train a small number of pairs of samples with significant sentiment, the classification performance of the method exceeded that of the full training samples. The SVM corpus classification achieves 81.89%, 85.17% and 83.97% for completeness, accuracy and F-measure, respectively. 83.87%, 84.76% and 85.12% for IG corpus classification, respectively. 83.65%, 84.12% and 82.89% for DF corpus classification, respectively. and 82.89%, respectively. In this paper, the detection rate, accuracy rate and F-measure rate of the hybrid method using semantic understanding and machine learning are 86.78%, 88.89% and 87.98%, respectively. In the annotated training text set, using the hybrid method of semantic understanding and machine learning to select the training samples can not only reduce the training time, but also improve the accuracy rate.

5. Conclusion

This paper firstly introduces the concept of Chinese text sentiment classification and the four steps of Chinese text sentiment classification by semantic understanding and the process of Chinese text sentiment classification by machine learning, which focuses on three classification models. We also introduce the data collection and the processing and labeling of the corpus for the implementation and application of the Chinese text sentiment classification algorithm framework, and then use the hybrid approach of semantic understanding and machine learning to design the classification model in detail. The accuracy of the classification is high.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Tubishat, Mohammad, Idris, et al. Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. Information Processing & Management: Libraries and Information Retrieval Systems and Communication Networks: An International Journal, 2018, 54(4):545-563. https://doi.org/10.1016/j.ipm.2018.03.008

[2] Lee, Pei-Ju, Ya-Han, et al. Assessing the helpfulness of online hotel reviews: A classification-based approach. Telematics and informatics, 2018, 35(2):436-445. https://doi.org/10.1016/j.tele.2018.01.001

[3] Tubishat, Mohammad, Idris, et al. Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. Information Processing & Management: Libraries and Information Retrieval Systems and Communication Networks: An International Journal, 2018, 54(4):545-563. https://doi.org/10.1016/j.ipm.2018.03.008

[4] Le C C, Prasad P, Alsadoon A, et al. Text Classification: Naieve Bayes Classifier with Sentiment Lexicon. IAENG Internaitonal journal of computer science, 2019, 46(2PT.141-263):141-148.

[5] Acharya, Vishwanath, Bora, et al. Classification of SDSS photometric data using machine learning on a cloud. Current Science: A Fortnightly Journal of Research, 2018, 115(2):249-257. https://doi.org/10.18520/cs/v115/i2/249-257

[6] Donya, Dezfooli, Seyed-Mohammad, et al. Classification of water quality status based on minimum quality parameters: application of machine learning techniques. Modeling Earth Systems and Environment, 2018, 4(1):311-324. https://doi.org/10.1007/s40808-017-0406-9

[7] Kolkur M S, Kalbande D R, Kharkar V. Machine Learning Approaches to Multi-Class Human Skin Disease Detection. International journal of computational intelligence research, 2018, 14(1):29-39.

[8] Dhar P, Dutta S, Das P, et al. Cross-wavelet aided ECG beat classification using LIBSVM. Computer Methods in Biomechanics & Bio, 2018, 6(3-4):343-352. https://doi.org/10.1080/21681163.2016.1251339

[9] Priyadarshiny, Dhar, Saibal, et al. Cross-wavelet aided ECG beat classification using LIBSVM. Computer methods in biomechanics and biomedical engineering. Imaging & visualization, 2018, 6(3-4):343-352. https://doi.org/10.1080/21681163.2016.1251339

[10] Vasundhara D N, Seetha M. Rough Set based SVM Technique for Spatial Image Classification. International Journal of Computational I, 2019, 14(1):27-40.

[11] Sampaio W, Oliveira F, Filho A, et al. Classification of breast tissues into mass and non-mass by means of the micro-genetic algorithm, phylogenetic trees, LBP and SVM. Computer Methods in Biomechanics & Bio, 2018, 6(3-4):315-330. https://doi.org/10.1080/21681163.2016.1240630

[12] Vala M. Document Classification: A Technical Review. National Journal of System and Information Technology, 2018, 11(1):67-74.

[13] Hamid Y, Journaux L, Lee J A, et al. A novel method for network intrusion detection based on nonlinear SNE and SVM. International journal of artificial intelligence and soft computing, 2018, 6(4):265-286. https://doi.org/10.1504/IJAISC.2018.097280

[14] Chaitanya, Anne, Avdesh, et al. Multiclass patent document classification. Artificial intelligence research, 2018, 7(1):1-14. https://doi.org/10.5430/air.v7n1p1

[15] Brindha N, Visalakshi P. Content Based Video Feature Extraction and Classification Using Perceived Motion Energy Spectrum-SVM Classifier. International Journal of Computing & Information Technology, 2018, 10(1):1-10.

[16] N, Brindha, P, et al. Content Based Video Feature Extraction and Classification Using Perceived Motion Energy Spectrum-SVM Classifier. International journal of computing & information technology, 2018, 10(1):1-10.

[17] Marrugo N, Amaya D, Ramos O. Comparison of Multi-Class Methods of Features Extraction and Classification to Recognize EEGs Related with the Imagination of Two Vowels. International Journal on Communications Antenna and Propagation, 2018, 8(5):398-405. https://doi.org/10.15866/irecap.v8i5.12709

[18] Pampoulou E, Fuller D R. Introduction of a new AAC symbol classification system: the multidimensional quaternary symbol continuum (MQSC). Journal of Enabling Technologies, 2021, 15(4):252-267. https://doi.org/10.1108/JET-04-2021-0024