Cognitive Computation (2021) 13:1114–1127
https://doi.org/10.1007/s12559-021-09855-4

# Aspect-Based Sentiment Analysis for User Reviews

**Jinyang Du[2] · Yin Zhang[1] · Xiao Ma[2] · Haoyu Wen[2] · Giancarlo Fortino[3]**

Received: 26 August 2020 / Accepted: 17 March 2021 / Published online: 13 July 2021


**Abstract**
Aspect-based sentiment analysis (ABSA) can help consumers provide clear and objective sentiment recommendations through massive quantities of data and is conducive to overcoming ambiguous human weaknesses in subjective judgments. However, the robustness and accuracy of existing sentiment analysis methods must still be improved. We first propose a deep-level semiself-help sentiment annotation system based on the bidirectional encoder representation from transformers (BERT) weakly supervised classifier to address this problem. Fine-grained annotation of restaurant reviews under 18 latitudes solves the problems of insufficient data and low label accuracy. On this basis, bagging traditional machine learning algorithms and annotation systems, a novel classification model for specific aspects is proposed to explore consumer behavior preferences, real consumer feelings, and whether they are willing to consume again. The proposed approach can effectively improve the accuracy of the ABSA tasks and reduce the space-time complexity. Moreover, the proposed model can significantly reduce thequantity of data annotation engineering required.

**Keywords** Aspect based · Sentiment analysis · Machine learning · Cognitive computing

## Introduction

In recent years, a lot comments posted by users in various applications on the internet often contain rich sentiment opinions behind them. In particular, a large number of user-generated comment texts appear on online catering platforms. In addition, with a sharp increase in public reviews, a large number of user-generated reviews have appeared on online catering platforms. These review data have significant potential and are a great reference for consumers and businesses. For consumers, the evaluations of other consumers help understand the quality of a particular business and are an important reference when choosing a restaurant. In addition, access to a large number of accurate and objective user reviews helps merchants improve their catering products.

✉ Yin Zhang
yin.zhang.cn@ieee.org

[1] University of Electronic Science and Technology of China, 4 E 2nd Section, 1st Ring Rd, Jianshe Road, Chenghua District, Chengdu, China

[2] Zhongnan University of Economics and Law, Wuhan, China

[3] University of Calabria, Via Pietro Bucci, Rende CS 87036 Arcavacata, Italy

However, due to the large number of online reviews, it is difficult to manually view and read all customer reviews to extract valuable information. Therefore, there is an urgent need to express information in a form that the computer can "understand" to evaluate and analyze the data to reduce labor costs. Sentiment analysis [1] precisely solves these problems. Sentiment analysis is one of the essential tasks and challenging problems in the field of natural language processing. Its purpose is to use text mining and other methods to identify and extract subjective information in the text to judge the sentiment polarity of the text (positive, neutral, negative, etc.), which is divided into three subtasks based on fine-grained research: text level, sentence level, and aspect level.

In general, the sentiment polarity of user reviews should be analyzed from multiple granularities. However, a traditional sentiment analysis often separates various aspects. It considers the sentiment of the text from only one aspect or a few dimensions. As a result, researchers have focused on analyzing sentiments from multiple granularities and improvements from existing algorithms. ABSA aims to predict the object of a sentiment expression and the polarity of the emotion [2, 3]. However, model training used in ABSA requires a large quantity of labeled data. The time and spatial overhead are high. Therefore, this paper intends to make full

use of smaller labeled data and analyze user evaluations in the catering industry from multiple granularities.

The ABSA studied herein aims to predict real sentiment tendency of user reviews and the willingness of the users to consume again. Concerning the application value, by using computer technology to refine and analyze many reviews and intelligently complete the classification of true emotions, this model will help consumers choose products and help merchants collect suggestions. This article also analyzes user evaluations in the food and beverage industry from multiple granularities, mining the implicit sentiment characteristics of consumers and their willingness to consume again, which significantly reduces the quantity of data annotation engineering required. The shortcomings of high time and spatial overhead improve the accuracy in distinguishing true emotions in ABSA. Specifically, the contributions of this paper are as follows:

- The present study on aspect-based sentiment annotation features aims to replace traditional text representation features, build a text feature annotation system for the catering industry, and breakthrough the bottleneck of single-dimensional sentiment classification.
- Based on a restaurant review dataset, four traditional machine learning techniques are used to build aspect-based sentiment classifiers.
- An ensemble aspect-based model is proposed to improve the accuracy of the sentiment classification results and reduce the space-time complexity.

## Related Work

### User Review Analysis

In terms of movie reviews, Pang and Lee [4] previously studied how to analyze user reviews and proposed using machine learning to mine sentiment tendencies. Using movie reviews on the Internet as a corpus, different feature selection methods were applied. Using naive bayes, maximum entropy, and a support vector machine (SVM) to classify movie reviews, the experimental results showed that the model achieves the best classification performance and the highest level of accuracy. Yang et al. [5] modeled the sentiment relevance of customer reviews as a graphical representation. Zhang et al. [6] proposed a trust representation model for sentiment similarity analysis.

For the field of online game reviews, the data are more complicated. The age of the game group is relatively younger. They use some professional gaming terms or online language. Hence, the text content is difficult to understand and difficult to handle. Traditional methods are inefficient in handling such comments. In this field, Stråat et al. [7] proposed a manual aspect-based sentiment analysis approach. Existing work [8, 9] only uses some lexical-level features such as word frequency. Although these methods combine traditional machine learning algorithms to judge commenting users' sentiment, they cannot mine sentiment concerns for specific aspects. Our method combines deep learning and traditional machine learning. It can use a weakly supervised semiself-help sentiment annotation system and bagging algorithm to fully judge the user's sentiment tendency on massive review data.

### Sentiment Analysis Approaches

Recently, researchers have conducted numerous studies [10, 11] on improving the accuracy and generalization of sentiment analysis models. The research on sentiment analysis based on a sentiment polarity can be traced back to 2000. For example, Huettner and Subasic [12] used a fuzzy logic method to classify documents. During the analysis process, a manually constructed sentiment polarity was used to classify sentiment words in detail, including semantic classification, part-of-speech stacking, and sentiment intensity quantification. Bravo-Marquez et al. [13] used two different sentiment dictionaries (Opinion Lexicon and Davar Lexicon) to conduct a sentiment analysis on Twitter. Then they used a scoring method to analyze the sentiment tendency of the text, including calculating the text word frequency-reverse document frequency (TF-IDF) [14]. Inspired by previous studies [15], Stevenson et al. [16] collected descriptive data on the Affective Norms for English Words (ANEW) to identify which discrete emotions are elicited by each word in the set. Cambria et al. [17] integrate top-down and bottom-up learning via an ensemble of symbolic and subsymbolic AI tools, which apply to the interesting problem of polarity detection in a text. The text and correct adjectives as the evaluated words were matched to strengthen the ability of the affective polarity and improve the accuracy. However, with the emergence of many online vocabulary and spoken words in review sentences, the existing studies cannot identify vocabulary with emotions, resulting in unsatisfactory classification results; thus, machine learning methods have been used for sentiment analysis. Traditional sentiment analysis methods include many mature models, such as the hidden Markov model (HMM) [18] and conditional random field (CRF) [19]. Among them, it can be seen from the deep neural network experiments conducted by Kadari et al. [20] that Long Short-Term Memory (LSTM) [21] achieves a better performance than CRF for sequence stacking tasks and is also better than other types of neural networks. Some studies have used attention-based methods for sentiment analysis, Basiri et al. [22] proposed a bidirectional cnn-rnn deep model with attention

mechanism to select the important features. El-Affendi et al. [23] proposed a novel deep learning-based multilevel attention neural (MPAN) model applied to perform Arabic sentiment analysis tasks. Akhtar et al. [24] proposed a stacked ensemble model for predicting the degree of intensity for emotion and sentiment. In addition, a gated recurrent unit (GRU) has been proven to achieve a similar performance as LSTM in numerous fields, despite its simpler structure [25].

### Aspect-Level Sentiment Analysis

However, the traditional polarity sentiment classification model ignores the sentiment polarity of each granularity, for example, "The food at this restaurant tastes very good, but the location is too far away." In the previous sentence, two review objectives were mentioned: taste and location. The emotions expressed for the first goal are positive, whereas those for the second are negative. In a traditional sentiment classification task, only the comprehensive sentiment polarity of the text is obtained. In ABSA [26, 27] field, extraction problems for the targets and classification problems for the target emotions occur. Identifying the right target and classifying the target sentiment correctly are important ABSA tasks. Aisopos et al. [28] determined the overall sentiment polarity of Weibo by comparing content and context features. They used context features to propose the concept of a sentiment polarity rate for the relationship between Weibo reviews. In terms of content features, an n-gram feature was proposed to improve the accuracy of ABSA. Tang et al. [29] proposed a memory network and attention mechanism to solve certain aspects of sentiment analysis. Unlike SVM, LSTM, and other models, this model can accurately capture the importance of each context word when inferring each aspect. This level of importance and a text representation are obtained through multiple layers of calculations, and each layer is based on an external memory attention neural network model. When analyzing the experimental results, not only was the effectiveness of the model proven, the model was also shown to be faster than other methods. Wang et al. [30] proposed a word-level and clause-level attention networks for ABSA. Li et al. [26] proposed predicting the sentiment polarity while conducting an extraction, and they converted the target extraction problem and sentiment classification problem into two serial sequence stacking problems. A unified stacking model was designed, based upon which a neural network model that solves the problem of a unified end-to-end multiaspect stacking model was proposed. Majumder et al. [31] used a transfer learning method, using a pretrained aspect extraction model to improve the effect of ABSA tasks. Dashtipour et al. [32] proposed a rule-based method using dependency grammar and deep learning for polarity detection in the Persian language. Ma et al. [33] proposed a two-step attention model that can utilize affective common sense to develop a common sense system for target aspect-based sentiment analysis.

Based on existing studies, the traditional text sentiment classification model judges the sentiment polarity from a certain aspect. However, there remains room for improvement in the accuracy of true sentiment discrimination. Among the sentiment analysis tasks based on traditional machine learning methods, machine learning algorithms have a limited ability to learn the text representation features, ignoring multimodel ensemble strategies with different characteristics and leading to a problem in which the accuracy of traditional machine learning sentiment classification models is insufficient. With the advantages regarding the efficiency of traditional machine learning models, the accuracy of their models can still be improved. However, the existing deep learning pretrained language models have achieved fairly great results in terms of the accuracy of emotion discrimination based on a single aspect, the training time and space required remain high. The model combining the advantages of both will have significant research value.

## Framework

Currently, researchers have achieved good results in ABSA. For example, Li et al. [26] considered the issues of aspect-based recognition and sentiment classification as two sequence stacking problems using two concatenations. A neural network model comprehensively solves the problem of ABSA with high time and spatial overhead. To quickly understand the user's real feelings regarding their consumption and willingness to consume again through machine learning and to improve the prediction accuracy of existing models, as well as the time and spatial complexity of language models, in this study, we built an ensemble stacking system for the food and beverage industry. In addition, we applied the efficiency of traditional machine learning models and the generalization ability of ensemble models to solve the application problems of ABSA. ABSA is constructed based on ensemble learning from the following aspects. 1) The multiaspect stacking model was built for restaurant reviews. 2) Four classifications were built based on traditional machine learning technology, namely, a granularity emotion classifier, which lays the foundation for realizing the ensemble model. 3) A multiaspect stacking model based on ensemble learning is proposed.

Our work focuses on the mobile Internet catering industry. It proposes a series of feature construction and model training methods based on the characteristics of online reviews on mobile terminals. This section mainly describes how to construct aspect-based features, how to integrate the modeling, and how to build the models based on the application scenarios.

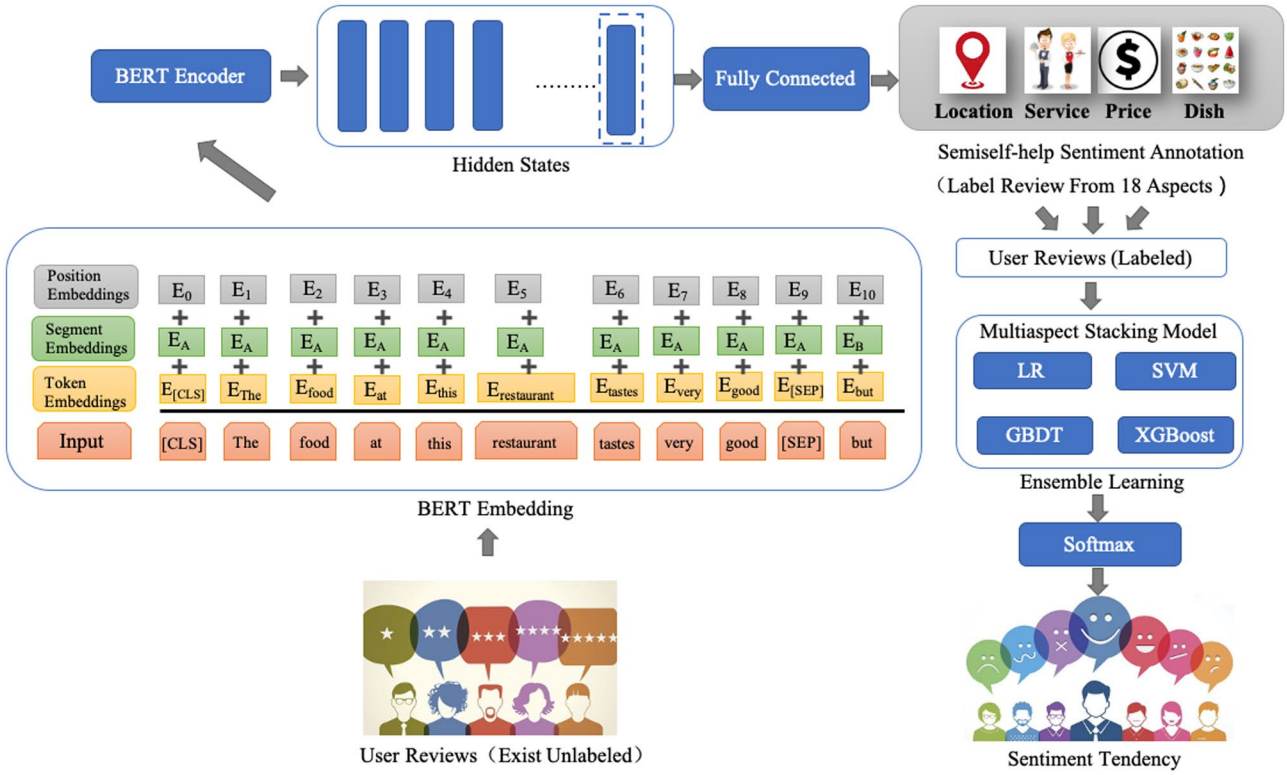


**Fig. 1** Framework of the model

## Task Definition

**Input**: User Reviews T

**Output**: Consumption feeling expressed in each review text and willingness to consume again.

In this study, ABSA is used as technical support for predicting the actual consumption feelings of users and their willingness to consume again. Statistics-based sentiment analysis tasks often require a large quantity of supervised text data to achieve accurate and robust model classification. However, when large-scale labeled datasets cannot be obtained, a powerful annotation system needs to be defined to enhance the dataset. The second step is to perform fine-grained sentiment analysis on 18 aspects and finally use the final ensemble learning result.

## Model Description

Based on the above tasks, the overall framework of this study is shown in Fig. 1. After obtaining the initial data, the multiaspect stacking model is built based on the research reports and comment data of service industry researchers. The effectiveness of single-model training is determined. A good single model is used as the base learner for the ensemble model.

- Multiaspect sentiment label

In many texts, each dimension has a sentimen evaluation, such as "Wow, this restaurant is too delicious, but the location is too difficult to find, and the environment needs to be improved." This text starts with taste and location. Three granularities of the environment have been evaluated. What is the consumer feeling? Will the writer want to come again? These issues are worth studying and are considered in this paper. Therefore, 5 major aspects based on restaurant reviews, 18 small levels, that is, multiaspect stacking, and feature construction are described herein. According to the research report of the service industry, scholars have research scholars proposed research on the consumption trends and marketing strategies of the catering industry: sufficient portions, low prices, novel features, service quality, health and safety, interior design, brand size, and parking spaces are all factors that consumers value. Therefore, this article combines many corpora in the field of catering with the research reports of research scholars. A multiaspect stacking model with 18 subdivisions is proposed.

As shown in Fig. 2, this article proposes 5 major levels and 18 aspects of the multiaspect stacking model for the restaurant industry. This system is used to label a small portion of the original data (provided by the AI challenge platform).

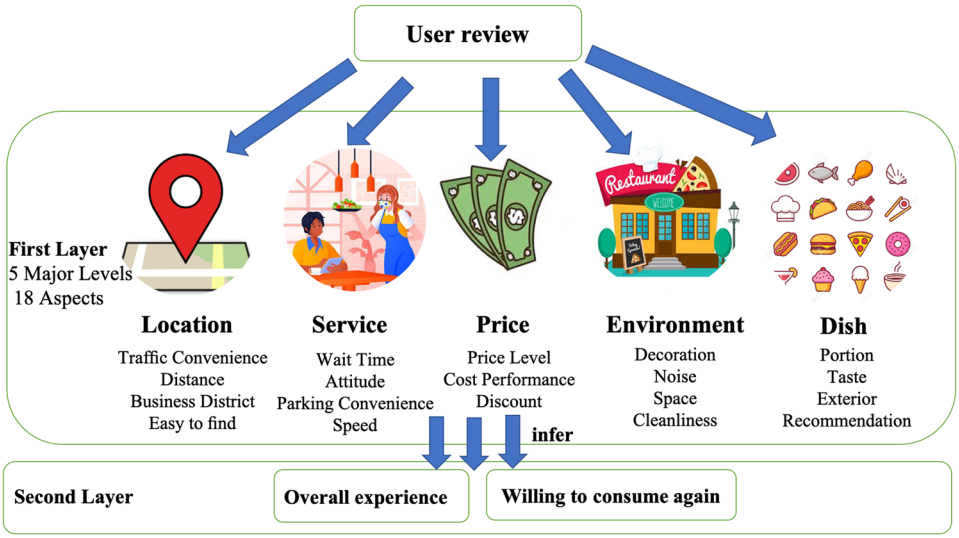


**Fig. 2** Restaurant review aspect

- BERT semiself-help sentiment annotation

After the multidimensional sentiment analysis method is proposed for a small quantity of data, the model requires fewer data and more generalized features. Even without many labeled text datasets, a good result can be obtained for unknown data. In this paper, a semiself-help multiaspect stacking model based on BERT [34] is designed to realize the automation of a multiaspect stacking model. This article introduces the following design ideas:

1. Use of cross-validation to denoise the data. Data samples are labeled by the supervisor when stacking; in particular, excessive subjective consciousness is involved, which can induce errors. In this study, cross-validation data are used to reduce noise, and the samples with a noise reduction greater than a certain thresholds are masked to achieve the denoising process. After this process, we can obtain a relatively complete and correct dataset.
2. After data validation, the text was input based on the character level. The calculation of input embeddings in BERT is carried out using three different embeddings. As shown in Fig. 1, it is computed by summing token, segment, and position embeddings. Token embedding is the vector representation of each token in the vocabulary. Position embeddings are used to preserve the information about the position of the words. Segment embeddings are used to distinguish between sentences. $< CLS >$ is a unique token of each sequence for classification, $< SEP >$ is the token separating sentence, Its input representation is constructed as:

$$h_i^0 = E_i^{tok} + E_i^{pos} + E_i^{seg} \tag{1}$$

where $E_i^{tok}$, $E_i^{pos}$, and $E_i^{seg}$ are as the corresponding token, segment, and position embeddings for $s_i$.

3. The input representation described above is then fed into L successive transformer encoder blocks, The BERT encoder is constructed by making use of transformer blocks. These blocks are used in 12 layers, each of which consists of 12 multi-head attention blocks. Input and output are aspect-based during pretraining, and shallow parameters are shared.

$$h_i^{\ell} = \text{Transformer}\left(h_i^{\ell-1}\right), \ell = 1, 2, \cdots, L \tag{2}$$

We denote the final hidden states as $\left\{h_i^L\right\}_{i=1}^N \in \mathbb{R}^{d_1}$, where $N = m + 1$. To obtain a fixed dimensional pooled representation of the input sequence, we use $h_0^L$ which is the first token $< CLS >$ of final hidden state.
4. The fully connected layer with the softmax cross-entropy loss classification function (1) minimize the cross-entropy value to optimize the network weight. where the real label y and the predicted sentiment label $\hat{y}$ are as follows:

$$\mathcal{L} = -\sum_i \sum_j y_i^j \log \hat{y}_i^j \tag{3}$$

where $i$ is the index of a data sample, $j$ is the index of a sentiment class.
5. Finally, label the dataset from 18 aspects.

BERT encompasses 11 natural language processing tasks through pretraining and aspect-based methods and incorporates a transformer encoding and decoding mechanism, which is more efficient than that used in recurrent neural networks and can capture long-range dependencies. Compared with other pretrained models, it captures real bidirectional context information. This paper uses BERT to train and learn from unknown expectations. Softmax is used downstream to classify sentiment objects, which contrasts strongly with traditional machine learning models.

### Single-Model Selection and Training and Improved Ensemble Bagging Algorithm

Based on traditional sentiment classification, in this study, a single-model algorithm is first used to directly model consumer consumption feelings and willingness to consume again. That is, four traditional machine learning models are applied: logistic regression (LR) [35], SVM, gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost) [36] algorithms. In this study, LR is used as the base model because it is simple to implement and is widely applied in industrial problems. The number of calculations is extremely small, the speed is fast, and the storage resources are low. In addition, it can be used to conveniently observe the sample probability score. For the case of linear inseparability, an SVM can be mapped to a high-dimensional feature space to achieve linear separability through the kernel function. The SVM learning problem can be expressed as a convex optimization problem, and thus a known effective algorithm can be used to find the global minimum of the objective function. Other classification methods (such as rule-based classifiers and artificial neural networks) use a greedy learning-based strategy to search the hypothesis space. In contrast, this method can generally only obtain a locally optimal solution. Tree model-based GBDT prediction achieves high accuracy and can handle nonlinear data. It can flexibly process various types of data, including continuous and discrete values. For a relatively small adjustment time, the prediction preparation rate can also be relatively high compared to an SVM. With some robust loss functions, it is extremely robust to outliers. In this study, XGBoost was used as a model to add the complexity of the tree model to the regular term and avoid overfitting, and thus, the generalization performance was better than that of the GBDT. The loss function is expanded through a Taylor expansion. The first and second derivatives are used simultaneously, which can speed up the optimization. In addition, GBDT only supports the classification and regression tree (CART) as the base classifier, and it also supports linear classifiers. When using linear classifiers, L1 and L2 regularization can be applied.

**Algorithm 1** Bagging multimodel algorithm

**Input:** Comment text dataset $D$, weak classifier algorithm $M_j$ Set $D = x_1, y_1, x_1, y_2 \ldots, x_m, y_m$ total documents, divided into n categories

**Output:** Classification results, evaluation index values

1: Determine $k$ and $T$ (where $k$ is the number of rounds for folding, and $T$ is the number of different weak classifiers)
2: **for** $i = 1 \rightarrow k$ **do**
3: **for** $j = 1 \rightarrow T$ **do** randomly return the sampling $m - \frac{m}{k}$ data from $D$ to $A_i$ use $A_i$ as the training set of the $M_j$ algorithm to obtain a weak classifier $C_i$
4: **end for**
5: **end for**
6: For the text sample to be classified, the category is determined by voting based on the values of $k \times T$ classifiers
7: Return the classification results

The single-model sentiment classification task aims to obtain the overall score of a product or store for review. The improved ensemble learning algorithm (bagging multimodel) can make the model more generalized and accurate. The specific steps are as follows:

We model aspect-based sentiment classification from the traditional machine learning model through the above algorithm steps and can directly derive polar sentiment classification. This uses LR algorithms based on decision boundaries, SVM algorithms, gradient boosted tree algorithms based on tree models, and extreme gradient boosted trees.

According to the model proposed above, different feature extractions are applied separately. In general, text data can be transformed into TF-IDF, word2vec [37], and word representation features based on language models applied in sentiment analysis methods. Based on the construction of a multiaspect stacking model, multiaspect features for restaurant reviews are introduced in this study.

TF-IDF is a commonly used text representation method for a traditional text sentiment analysis. It can be used to extract the frequency of each word in a sentence and the frequency of an inverse document. Although the word frequency and sentiment classification correlation are not high, it can express text information quickly and easily. Word2vec is a hidden layer parameter model generated during the task of predicting context words using neural network models. It has high-dimensional characteristics and can more truly represent the true meaning of words. For deep learning language models, the text representation of the pretrained language model is also a hidden layer parameter. Compared with word2vec, it can not only learn the semantics of a text expression, but also obtain the characteristics of a short syntax level, subject-predicate level, and sentence structure level. The multiaspect feature can represent a series of factors strongly related to sentiment polarity. It can improve the accuracy of the model for sentiment classification tasks.

TF-IDF is a common textual representation of traditional machine learning models. It is simple and easy to obtain and can be used to quickly build benchmark models. Word2vec, to a certain extent, acts as an extended feature and can more accurately represent the true meaning of the text. As a traditional machine learning model, in this study, the word2vec mean method is used as the text feature representation, which can effectively improve the accuracy. The word representation based on BERT can learn higher-level semantic representations and use a two-way encoding and decoding mechanism to focus on the context relationship of the sentence, reducing the errors caused by a polysemy. Based on the features of the aspect-based system proposed in this paper, representative sentiment characteristics are applied. Using highly correlated features, the model can achieve a better prediction. When reviewing the

feature representation of a text, the text is long and the number of dimensions of the words is large. One-hot encoding causes significant redundancy and sparseness of the matrix, and thus the word2vec word frequency representation is used.

### Multiaspect Stacking Model

In the multiaspect stacking model, the base learners are all based on machine learning algorithms. In the field of restaurant reviews, this article summarizes 5 major aspects and 18 levels of detail. From these 18 levels of detail, an analysis model of the final polar emotion is developed. Considering that the learner built using only one machine learning algorithm will have insufficient diversity, this paper uses four different algorithms: LR, SVM, GBDT, and XGBoost, based on the ensemble stacking algorithm.

In this section, the principles and advantages of different ensemble algorithms are introduced. In practice, because of the possibility of repeated learning of samples in ensemble learning, it is necessary to improve the independence between models as much as possible when designing an ensemble method.

It reduces any overfitting that may occur, whereas, on the other hand, it reduces the possibility of "all things being wrong." In this paper, the following two principles are proposed for designing an ensemble method:

- The difference in the training set between individual learners is as large as possible.This principle is easy to understand. When the original training set of the model is different, no correlation problem occurs. The optimal training set between individual learners is a completely independent and identically distributed dataset. The principle that the training set of the individual learner must be as large as possible is extremely easy to follow for online advertising prediction with massive data.
- The differences in the model structure between individual learners are as large as possible. Because the model is used to simulate real situations, its structure will also cause errors in the prediction results. Such errors are often called "structural losses". Assume the structural differences of the model between individual learners are kept as independent as possible. In that case, the loss caused by the same type of model can be reduced. Few models have a positive effect on the prediction of highly sparse samples. In this study, an appropriate model is chosen as the base model of the ensemble model. Machine learning models can be roughly divided into three categories: generalized linear models, tree models, and network models. A network model is an ensemble model, the idea of which was originally inspired by a neural network. However, owing to the poor efficiency of the model, this paper mainly selects a suitable model from a generalized linear model and tree model as the base model to improve the structural difference between individual learners as much as possible. To fully utilize the characteristics of different models, the ensemble model designed in this study contains four individual learners with very different structures. These individual learners are themselves ensemble models and are homogeneous learners of tree models and generalized linear models. In addition, a heterogeneous learner is composed of a tree model and a generalized linear model. The GBDT has a fast learning speed and can efficiently process continuous features. However, it is still prone to overfitting when it has too many iterations. Therefore, when designing the ensemble model in this study, the number of GBDT iterations is deliberately controlled, multiple weak GBDT is trained as the base learner and integrated into a strong learner with strong predictive performance through the stacking method. A strong learner is an individual learner in a complete learner.

Different models have different advantages because of the nature of the model. The choice of a weak learner directly affects the generalization ability and accuracy of the ensemble model. As a nonlinear model, the tree model can ideally deal with continuous features; it can spontaneously find various distinguishing feature combinations (the path from the root of the tree to a certain leaf node can be regarded as a feature combination). The generalized linear model makes it easy to perform parallel computing and is good at dealing with massive numbers of discrete features. Additionally, however, using a single model to simulate the true distribution of the samples will also cause systematic errors owing to the nature of the model.

However, the tree model is not good at processing absolute high-dimensional features. When leaf node division is applied, it easily causes only a few strong features to have the opportunity to be divided into features. If all features are considered each time, the performance will be poor. The tree model is essentially a mechanical memory of the sample features, which easily leads to overfitting.

The shortcomings of generalized linear models cannot be directly dealt with using continuous features. For example, the LR model will first use a sigmoid function to process continuous features or convert continuous features into discrete features based on experience, causing a certain degree of feature loss and distortion; introducing nonlinear parity factors requiring manual feature crossing, which easily explodes, is extremely labor intensive and cannot exhaust all feature combinations.

Therefore, in this study, the concept of ensemble learning is applied. Two generalized linear models are used, i.e., a model based on the decision boundaries and two tree models as the base learner of the ensemble model. XGBoost and GBDT from the tree model family, with a better generalization performance, and LR and an SVM from the generalized linear model family, are used as the base model. The serial ensemble model with a better prediction performance through serial training is called a "multiaspect stacking

model" as a complete learner. The principle and structure of the stacking algorithm are as follows.

The algorithm flow and pseudocode are as follows:

**Algorithm 2** Multiaspect stacking model algorithm

**Input:**

Training set: $X \in R$ (Contains $N$ samples)
Base model: $Mi = LR, SVM, GBDT, XGBoost$
Secondary learner: XGBoost
Base model iterations: $d$
Training set partition: $k$

**Output:**

Ensemble model: Stacking_M
$M = m||initializeStacking\_M$
$D_1, D_2, D_3, \ldots D_k = cut(D, k)$
**for** $i = 1 \rightarrow k$ **do**
    Gi=train($Mi, D.drop(Di)$)
**end for**
**for** $i = 1 \rightarrow k$ **do**
    Fi=predict($Gi, D$)
**end for**
XGB=train(XGB,F )
Stacking_M=Stacking(LR,SVM,GBDT,XGBoost,XGB)
Return Stacking_M

1. For LR, the training set D is divided into K parts. For each part, the model is trained using the remaining dataset, and the result of this part is then predicted.
2. The above steps are repeated until each part is predicted to obtain the training set of the secondary model.
3. K test sets, and the test set of the secondary model after averaging are obtained.
4. For SVM, GBDT, and XGBoost, the above situation is repeated, and the four primary models obtain four-dimensional feature data.
5. The secondary model is selected for training prediction, and the last layer uses the XGBoost algorithm.

# Experiments

## Experimental Design

### Experimental Ideas

Based on the modeling ideas described, to verify the sentiment analysis model of ensemble learning based on the multiaspect stacking model proposed in this paper, this section describes the designs of three experiments conducted to verify the feasibility and performance of the model:

- Verify the efficiency and model accuracy of the sentiment classification model based on the combination of traditional ensemble learning algorithms with improved bagging ideas and TF-IDF and word2vec features.
- Verify whether the accuracy of the weakly supervised BERT is sufficiently accurate for a single aspect.
- Verify that the ensemble model based on a multiaspect stacking model can achieve a high precision and low complexity performance.

Therefore, in this section, the empirical evidence of the traditional machine learning model and the pretrained language model BERT are used to obtain the experimental results as a baseline. The ensemble multiaspect stacking model is then applied. The ensemble multiaspect stacking model studied herein requires two polar sentiment classification models as technical support. Only the ability of traditional machine learning models can be improved to a certain degree. The training speed is faster than that of the language model. In addition, the accuracy of deep learning models for a single aspect is sufficiently accurate to combine their respective advantages and demonstrate an ensemble aspect-based model.

### Experimental Purpose

From the experimental design perspective, the feasibility of the model is verified herein based on three aspects.

- Through a demonstration of traditional machine learning models and the use of an improved ensemble algorithm, the model's efficiency and generalization capability are improved to allow its prediction results to achieve a higher score, that is, to verify that it is more efficient and has the ability to improve the accuracy of the results. Groundbreaking proof of its subsequent aspect-based ensemble approach is provided.
- BERT is used to classify the sentiment of the text, verify the high generalization ability of the language model and the high accuracy performance for a single dimension, and make an ensemble semiself-help multi-aspect stacking model.
- Based on the two empirical results above and considering various factors, the improved ensemble model is used to verify the high accuracy and efficiency of the divided test set. After semiself-help stacking of the original test set, the generalization of the model is verified. To prove the high accuracy and efficiency of the system integration modularity from numerous aspects, the model is proven to be feasible and applicable.

### Dataset

Based on the modeling ideas described, the initial data obtained from the AI challenge competition dataset include mobile online comment data and traditional sentiment classification tags. The purpose of the competition is to build an accurate model for predicting sentiment classification categories. The evaluation of the data based on the micro-F1-score is taken from the AI challenge platform. The section introduces the model design ideas and methods. The labeled supervision data released by the AI challenge platform are described for empirical analysis. Considering that the model is more generalized and accurate, the training data used in the text data classification problem cannot be too small, and thus 100,000 empirical data points are applied. The labeled review data are used as the training set, and 20,000 data points are used as the verification test set. After constructing the stacking system, the laboratory team collectively spent approximately one month completing the stacking of 23,000 aspects.

### Model Evaluation Index

There are many indicators used in machine learning classification to evaluate the model capabilities and achieve accurate results. In a binary classification problem, the performance index of the classifier is usually judged based on the accuracy, precision, recall, and F1 value of the classifier.

The recall rate is a measure of the recognition of the positive examples by the classifier, that is, how many actual positive examples can be accurately classified as positive examples by the classifier's prediction. It is worth mentioning that the sentiment analysis task is usually a three-category task. The categories are positive, neutral, or negative, or more categories are applied. At this time, F1 has two calculation methods. First, the total value of each category is calculated according to Eq. 4:

$$MicroF1 = \frac{2TP_{total}}{2TP_{total} + FP_{total} + FN_{total}} \tag{4}$$

For the second type, F1 first calculates the total value of each category, where N is the number of emotion categories, and the F1 value is then calculated according to $MacroF1 = \frac{1}{N}\sum_{i=1}^{N} F1^{i}$.

In the multiclassification problem of a sentiment classification described in this paper, as with the multiclass F1 value calculation method shown, it is also applicable for mixing the matrix in the multiclassification problem to the different classes, class K has K dimensions. Thus, when calculating, focus on the true, false positive, true, and negative values. The correct calculation can also be used to obtain the recall rate. The F1-score and other calculation indicators apply macro, micro, and weighted F1 values for result evaluation and comparison. In this paper, the MicroF1 value is used to evaluate and compare multiple classification models.

## Demonstration of the Single Model and Improved Bagging Model

### Data Acquisition and Preprocessing

The training set consists of 10,000 texts with supervised labels, and the test set contains 20,000 texts. Among them, the only training data are used as the content, which is the training text, and the tags are consumer feelings. The tags for willingness to consume again are (1, 0, -1), which represent positive emotions, neutral emotions, and negative emotions. The prediction target is a sentiment classification for 20,000 unlabeled data. As mentioned in the modeling in the previous section, the data preprocessing technology and the experimental process of data preprocessing are used to detect abnormal missing data from the dataset. In the process of determining the missing data, this study adopts the method of direct discarding because it is difficult to fill the text data with various filling methods. Regarding the abnormal data, in this dataset, we analyzed and detected garbled text and strange symbols. To avoid influencing the effect of the model, we also directly discarded such text. A total of 612 garbled data points and 387 abnormal symbol data points are removed. In fact, if this part of the data is retained, it will still have an impact on the results because the models cannot learn their useful information, and they do not have any valuable information about the review text in the catering field and should be removed according to the principle of information entropy. After removing these completely unusable data, this article uses a regularization method to cleanse and purify the data. Owing to different typing and writing styles of consumers, they will have many different symbols contained in the text, such as ("### #OW≫≫≫ It's delicious MMM"); in addition, there is too much useless information such as ("#", "≫") and other symbols in the middle, thereby increasing the complexity of the model training command. Inaccurate models may even allow the model to learn useless information, reducing the ability of the model. Therefore, in the preprocessing of this study, a regularization method is adopted to remove useless symbols contained in the text. We know that two symbols "!" and "?" are extremely meaningful when we express emotions, and thus we choose to keep them when dealing with text. The regular expression algorithm is as follows.

**Algorithm 3** Regular match removal algorithm

**Input:** Text: Reviews
**Output:** Cleansing Text: cleaned reviews
1: Determine $k$ and $T$ (where $k$ is the number of rounds for folding, and T is the number of different weak classifiers)
2: **for** each $i \in$ *Reviews* **do**
3: $i$-> drop special symbol
4: Clean Reviews-> append(i)
5: **end for**
6: Return cleaned reviews

After obtaining the cleaned text, because the smallest dimension of the Chinese text is a phrase or word, this article uses Jieba Chinese word segmentation to segment the text. The granularity of the tokenizer can be adjusted. Now, most tokenizers use some data to train a model. The model selected is generally a linear model, such as a CRF linear chain model. The purpose of word segmentation is to obtain the word vector based on the dimensions of the word, the flowchart of which is as follows in Fig. 3.

### Feature Extraction

Based on a traditional machine learning classification algorithm, its features are nongrid parameters, and each sample is a row in the feature matrix. Therefore, TF-IDF, countvectorizer, and word2vec are used to construct the feature matrix. Each text is converted into a combination of word representations or input features. CountVectorizer belongs to the common feature numerical calculation class and is a text feature extraction method. Each training text only considers how often each word appears in the training text. TF-IDF is used to evaluate the importance of words to a file set or a file in a corpus. The importance of a word increases in proportion to the number of times it appears in the file but, at the same time, decreases inversely with the frequency of its appearance in the corpus. In addition, word2vec is more complicated than the others and is a word vector obtained using neural network training. Each word has a corresponding vector. In general, if we consider a problem such as a comment sentiment analysis, we can directly average each word vector as a sentence vector when looking for a comment vector. This article demonstrates the first task of this study, which is designed using four single models and an ensemble model. The experiment uses three features to be input into the same classification-based algorithm to verify that the best feature representation is used as the model input. As shown in Fig. 4, regardless of what kind of base model is used, word2vec obtains the best results. From the experimental results, it can be concluded that the effect of the tree model is better than that of the generalized linear model, but this is from the base model, and there is no cross validation. This is simply used to determine the choice of the features. We can see that when selecting the TF-IDF, the results of several baseline models are not very different, as is the case with word2vec, although the word2vec feature can obtain a higher score, indicating that this feature is more stable. It can better match the model, and subsequent TF-IDF experiments need to be conducted.

### Analysis of Model Evaluation Results

At the end of this article, we describe the optimized parameters for every single model. Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm training is added to the LR baseline model and is the most commonly used method for solving unconstrained nonlinear programming problems. It has the advantages of fast convergence and low memory overhead and is often found in different machine learning algorithms. Specifically, L-BFGS has the same effect as gradient descent and SGD, although in most cases, it converges faster, which is important in large-scale calculations. The SVM introduces training of the Gaussian kernel function. The sampling ratio of the training set is 80%; GBDT and XGBoost are trained using cross-validation methods to obtain iterations. The learning rate, tree depth, width, and other parameters are set using a grid optimization algorithm. The grid step size is set to 1. Except for a set of optimal parameters, the ensemble strategy uses a weighted summation, and the learner weight is the prediction accuracy of each classifier on the validation set.

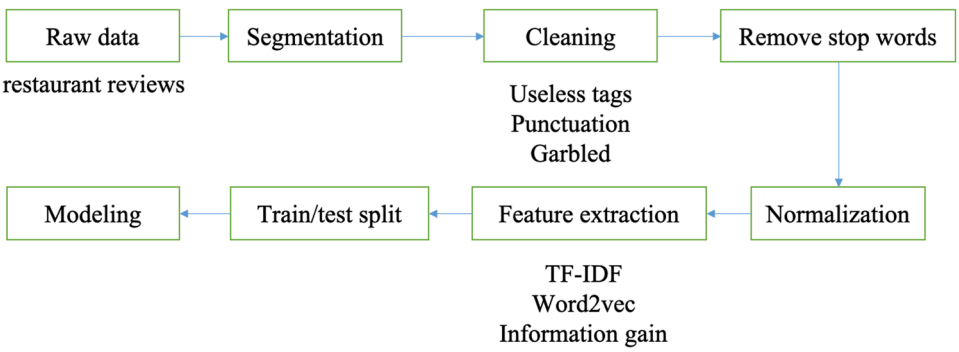


**Fig. 3** Data preprocessing flowchart

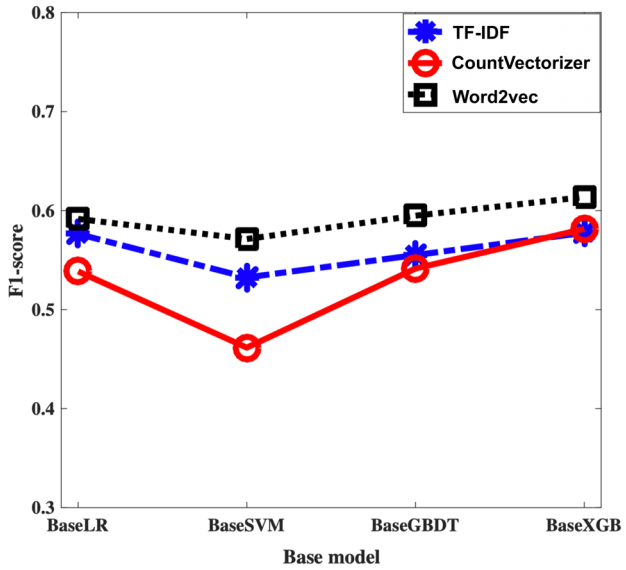


**Fig. 4** Model comparison

Finally, based on a traditional sentimental machine learning classification algorithm, the idea of bagging is introduced herein. The original bagging algorithm uses the same basic algorithm model and learns different datasets. In addition, different base learners are generated. In this study, the training set is divided into four parts, each using a 50% bagging voting mechanism, and four parts are finally integrated into the voting to obtain a bagging multimodel. This ensemble model includes different learners, namely, a total of $4 \times 5 = 20$ weak learners. The final result is determined based on the number of votes. In the experimental results, it was found that a bagging multimodel combining multiple algorithms is indeed better than single-model bagging. The recall rate is also higher. A comparison of its F1-score and recall rate is shown in Table 1.

In the final results, we can see that an ensemble model based on bagging can obtain a better score, reaching 0.6945. Its F1 score is also the highest among these models at 0.6568. The characteristics are obtained by the first section of verification. For the word2vec vector features, although the training set used the same set of 100,000 data points, the test set applied 20,000 data points.

**Table 1** Comparison of traditional machine learning sentiment classification models

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| LR(10w)-Word2vec | 0.5668 | 0.6631 | 0.6112 |
| SVM(10w)-Word2vec | 0.5425 | 0.6237 | 0.5803 |
| MLP(10w)-Word2vec | 05522 | 0.6645 | 0.6032 |
| Bi-LSTM(10w)-Word2vec | 0.5637 | 0.6705 | 0.6125 |
| GBDT(10w)-Word2vec | 0.5648 | 0.6712 | 0.6134 |
| XGB(10w)-Word2vec | 0.6228 | 0.6521 | 0.6371 |
| Bagging multimodel | 0.6230 | 0.6945 | 0.6568 |

## Evidence of BERT

In this experiment, a traditional polar sentiment classification of emotions is conducted using text reviews; that is, text review information is used to predict consumer feelings and willingness to consume again directly, and using a dataset including 35,000 user reviews, empirical evidence of single-grain accuracy is shown. Owing to the limitations of the experimental environment, it is infeasible to retrain the catering-oriented language model based on significantly large expectations. In this study, BERT, which is a Chinese model with 12 layers of encoders, 768 hidden layer nodes, and 12 multiheaded attentions, is applied. Next, the input and output are aspect-based, and softmax is added to the last layer of the network to achieve classification. BERT is used because it is based on extremely high expectations and is trained using a complex network transformer. After loading, we can directly initialize the shallow parameters of the pretrained model, that is, there is no need to spend more time or space learning information. In addition, the way the model handles text is different than described earlier and is based on word vectors, which are based on characters.

The total number of training parameters is over 100 million, which incurs high time and spatial requirements for the program to run. This study is based on the above experimental environment model, which lasted until the 22 h run was complete. Satisfactory results were also obtained, as shown in Table 2.

BERT has a higher score based on the 35,000 training data and the same test set owing to the environmental constraints and space-time complexity. This proves that the weakly supervised BERT has a large number of corpus words representing information. Compared with the other models, it can greatly improve the results.

**Table 2** Comparison of Consumption Willingness and Reconsumption Experience Score (BERT)

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| LR(10w)-Word2vec | 0.5668 | 0.6631 | 0.6112 |
| SVM(10w)-Word2vec | 0.5425 | 0.6237 | 0.5803 |
| MLP(10w)-Word2vec | 05522 | 0.6645 | 0.6032 |
| Bi-LSTM(10w)-Word2vec | 0.5637 | 0.6705 | 0.6125 |
| GBDT(10w)-Word2vec | 0.5648 | 0.6712 | 0.6134 |
| XGB(10w)-Word2vec | 0.6228 | 0.6521 | 0.6371 |
| Bagging Multimodel | 0.6230 | 0.6845 | 0.6568 |
| BERT(3.2 W) | 0.6866 | 0.7028 | 0.6946 |

**Table 3** Comparison of the ensemble multi-aspect stacking model

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| Bagging multimodel(original test set) | 0.6230 | 0.6845 | 0.6568 |
| BERT(3.2 W) | 0.6866 | 0.7028 | 0.6946 |
| Multiaspect stacking model(3.2 W) | 0.7968 | 0.8380 | 0.8169 |
| ElMo-LSTM(original test set) Blending(original test set) | 0.7044 | 0.7348 | 0.7193 |
| Multiaspect stacking model (semiautomatic annotation-original test set) | 0.7124 | 0.7412 | 0.7265 |

In a single granularity demonstration, the accuracy of the BERT for semiself-service stacking can reach 96.78%, and the F1 value can reach 0.8634. This also proves the feasibility of single-grained sequence stacking of the language model because it can use a large quantity of unsupervised data to learn the representation information of the language. This provides powerful technical support for the generalized model.

## Empirical Analysis of the Ensemble Multiaspect Stacking Model

In the empirical analysis described in the previous section, we demonstrated the strong learning ability of the BERT weakly supervised learning approach. It can be stated that this method achieves better results than the traditional machine learning emotion classification model. Even if an ensemble multiaspect stacking model is applied, the approach can still achieve better results than bagging. The model achieves a 4% higher quantile. The 4% improvement based on the large dataset applied is already quite significant. The ensemble multiaspect stacking model proposed in this study aims to apply an aspect-based and multidimensional method to achieve this improvement based on a particular catering area. A total of 18 highly relevant dimensions are used: traffic convenience, distance, business street, easy to finding, wait time, attitude, parking convenience, speed, price level, cost performance, discount, decoration, noise, space, cleanliness, portion, taste, exterior, and recommendation. The 18-dimensional system labels are applied as a feature input.

Finally, the model achieved good results on its own divided test set, and its F1-score reached 0.8169. Of course, this is based on a test set divided by the stacking system, which also had great results. However, it cannot be compared with other models described in the previous section because their test sets are different, although a score of 0.8169 is sufficient to prove the superiority of the model. However, to be more rigorous, this paper uses weakly supervised learning to predict the original test set of the 18 dimensions automatically. This experiment showed that the weakly supervised learning model can accurately learn the text classification results in terms of a subdivision at 96.78%. This result motivated the continuance of this research. BERT is multidimensionally labeled as the input feature of the test set. Based on this, the ensemble multiaspect stacking model is used to predict the test set.

The results are shown in Table 3 above, where '(3.2 W)' represents 32,000 training sets. After a round of semiautomatic stacking, the ensemble multiaspect stacking model can still achieve a great performance, with an F1 value reaching 0.7265. It can be stated that with the help of BERT, the model achieves extremely good results. Using the ensemble multi-aspect stacking model, its final effect surpassed the method developed by Cheng et al.[1]

This improved model is the Jieba word segmentation. The pre-training model uses embeddings from language models (ELMo), applying LSTM to obtain high-level features for a classification result of 0.7193, as shown in Table 3 above. In this study, the F1 value of the multiaspect stacking model based on the catering industry is 0.73% higher than that of Cheng et al.'s approach. In addition, its model training speed is much faster than that of the language model. A comparison of the efficiency is shown in the following Table 4.

The training of the ensemble multiaspect stacking model only takes 0.13 h, which solves the problem of the high time complexity of a language model. The ensemble multi-aspect stacking model achieves higher accuracy and a faster model training rate, which proves its greater effectiveness and significant applicability.

In this study, the importance of the characteristics of each dimension from multiple aspects of the model training was also derived. It is still in line with the concept of modern consumers. Empirical evidence shows that consumers

**Table 4** Comparison of Consumption Willingness and Re-consumption Experience Score (BERT)

| Model | Average training time |
|---|---|
| BERT(35000) (original test set) | 8.1 hours |
| Bagging multimodel | 0.8 hours |
| Multiaspect stacking model | 0.13 hours |

[1] https://mp.weixin.qq.com/s/W0PhbE8149nD3Venmy33tw

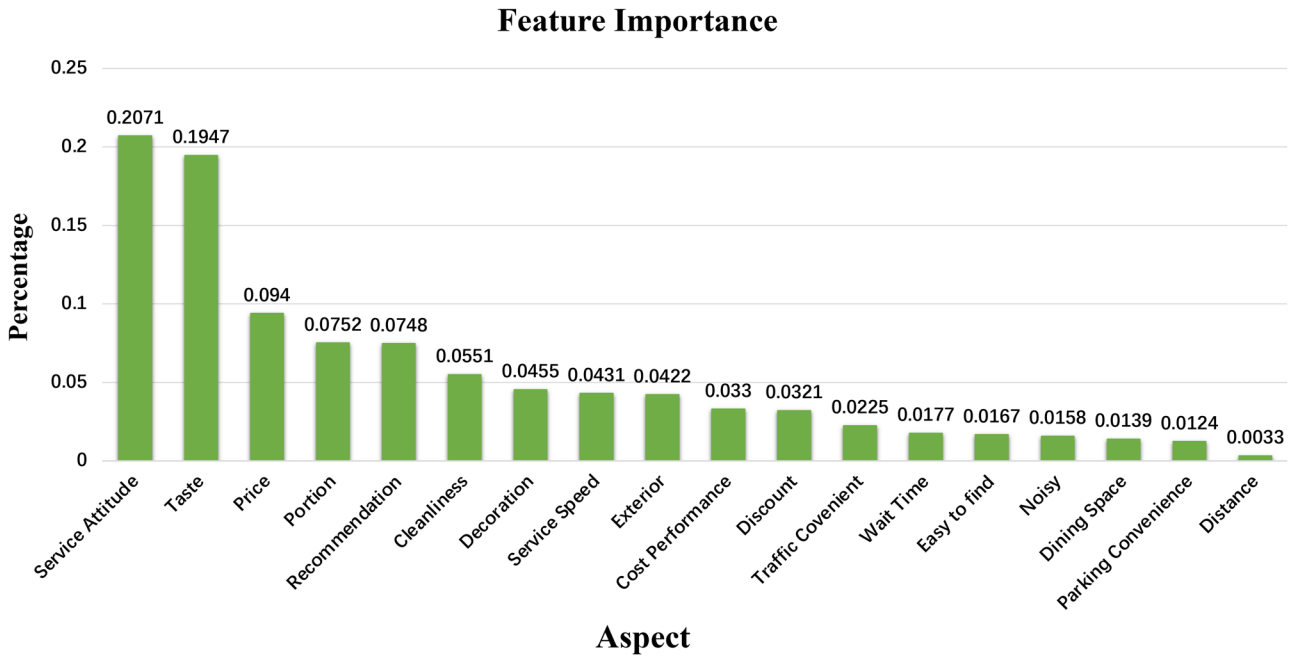


**Fig. 5** Feature comparison

tend to visit restaurants with good service and delicious food. Service attitude affects the consumption experience of the consumer more than taste, as well as the willingness to return. This study shows that if merchants want to obtain repeat customers, the quality of service should come first, followed by the taste of the food and the price.

It can be seen from the experiments that although the pretrained language model is worth studying, it is relatively difficult to apply to a specific field because based on the information from pure text, information unique to the field may not be learned but compared it with detail. In addition, a combination with a granular sentiment annotation system can obtain unexpected results. The sentiment analysis applied to the field of catering was substantiated herein.

## Discussion

As shown in Fig. 5, we sorted the importance of each dimension feature. We can see that it is very consistent with the consumption concept of modern consumers. In the top 25%, consumers are more inclined to choose the right service attitude and great taste. The results also show that to obtain repeat customers, businesses should strive to improve the quality of service and focus on whether the taste of catering is popular. 25% to 50% of the aspects continue to extend around service and food, including service speed, sanitary conditions, portion size, etc. The second half of the 50% aspect focuses on some additional factors surrounding the restaurant.

## Conclusion

This paper described the shortcomings of previous sentiment analysis methods, followed by existing text processing technologies, sentiment classification algorithms, language models, and ensemble learning theories. We proposed an improved bagging model, which is an ensemble model based on four classification algorithms that achieves enhanced ensemble bagging. It is also combined with a traditional sentiment classification algorithm and uses the BERT to construct a semiself-help multiaspect stacking model. The approach described herein is based on the two models; the multiaspect stacking model based on an aspect-based sentiment annotation system is used to optimize the accuracy of traditional machine learning models and improve the space-time complexity of existing language models. With the help of the BERT weakly supervised semiself-help multiaspect stacking model, the multiaspect stacking model is feasible and can obtain higher accuracy and better generalization.

However, owing to a lack of data, much work remains to be performed on the ABSA described in this paper. The model has fewer data in terms of a self-constructed stacking system, and there is still much room for improvement in tasks facing a single domain. In the future, we hope to expand the dataset using an ABSA system as much as possible to improve the capability of the language model and thus improve the ability to predict the actual consumption feelings of consumers and their willingness to consume again.

**Funding** This work was supported by the National Key R&D Program of China (No. 2020YFB1006002).

## Declarations

**Ethical Approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Conflict of Interest** Jinyang Du, Yin Zhang, Xiao Ma, Haoyu Wen and Giancarlo Fortino declare that they have no conflicts of interest.

## References

1. Cambria E. Affective computing and sentiment analysis. IEEE Intell Syst. 2016;31(2):102–7.
2. Akhtar MS, Ekbal A, Bhattacharyya P. Aspect based sentiment analysis: category detection and sentiment classification for hindi. In International Conference on Intelligent Text Processing and Computational Linguistics. Springer 2016. pp. 246–257.
3. Liu B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. 2012;5(1):1–167.
4. Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. 2008. 2(1–2):1–135.
5. Yang Q, Rao Y, Xie H, Wang J, Wang FL, Chan WH, Cambria EC. Segment-level joint topic-sentiment model for online review analysis. IEEE Intell Syst. 2019;34(1):43–50.
6. Zhang S, Zhong H. Mining users trust from e-commerce reviews based on sentiment similarity analysis. IEEE Access. 2019;7:13523–35.
7. Strååt B, Verhagen H, Warpefelt H. Probing user opinions in an indirect way: an aspect based sentiment analysis of game reviews. In Proceedings of the 21st International Academic Mindtrek Conference. 2017. pp. 1–7.
8. Strååt B, Verhagen H. Using user created game reviews for sentiment analysis: A method for researching user attitudes. In GHITALY@ CHItaly. 2017.
9. Zhu M, Fang X. A lexical analysis of nouns and adjectives from online game reviews. In International Conference on Human-Computer Interaction. Springer 2015. pp. 670–680.
10. Cambria E, Poria S, Hazarika D, Kwok K. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence. 2018. volume 32.
11. Huang F, Wei K, Weng J, Li Z. Attention-based modality-gated networks for image-text sentiment analysis. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2020. 16(3):1–19.
12. Huettner A, Subasic P. Fuzzy typing for document management. ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes. 2000. pp. 26–27.
13. Bravo-Marquez F, Frank E, Pfahringer B. Building a twitter opinion lexicon from automatically-annotated tweets. Knowl Based Syst. 2016;108:65–78.
14. Papineni K. Why inverse document frequency? In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Assoc Comput Linguist. 2001. pp. 1–8.
15. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett. 1999;9(3):293–300.
16. Stevenson RA, Mikels JA, James TW. Characterization of the affective norms for english words by discrete emotional categories. Behav Res Methods. 2007;39(4):1020–4.
17. Cambria E, Li Y, Xing FZ, Poria S, Kwok K. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020. pp. 105–114.
18. Lunter G. Hmmoc-a compiler for hidden markov models. Bioinformatics. 2007;23(18):2485–7.
19. Sarawagi S, Cohen WW. Semi-markov conditional random fields for information extraction. In Advances in Neural Information Processing Systems. 2005. pp. 1185–1192.
20. Kadari R, Zhang Y, Zhang W, Liu T. Ccg supertagging via bidirectional lstm-crf neural architecture. Neurocomputing. 2018;283:31–7.
21. Sak H, Senior AW, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.
22. Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR. Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. Future Gener Comput Syst. 2021;115:279–94.
23. El-Affendi MA, Alrajhi K, Hussain A. A novel deep learning-based multilevel parallel attention neural (mpan) model for multidomain arabic sentiment analysis. IEEE Access. 2021;9:7508–18.
24. Akhtar MS, Ekbal A, Cambria E. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]. IEEE Comput Intell Mag. 2020;15(1):64–75.
25. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint 2014. arXiv:1412.3555.
26. Li X, Bing L, Li P, Lam W. A unified model for opinion target extraction and target sentiment prediction. Proc Conf AAAI Artif Intell. 2019;33:6714–21.
27. Li X, Bing L, Zhang W, Lam W. Exploiting bert for end-to-end aspect-based sentiment analysis. arXiv preprint 2019. arXiv:1910.00883.
28. Aisopos F, Papadakis G, Tserpes K, Varvarigou T. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In Proceedings of the 23rd ACM conference on Hypertext and social media. 2012. pp. 187–196.
29. Tang J, Lu Z, Su J, Ge Y, Song L, Sun L, Luo J. Progressive self-supervised attention learning for aspect-level sentiment analysis. arXiv preprint 2019. arXiv:1906.01213.
30. Wang J, Li J, Li S, Kang Y, Zhang M, Si L, Zhou G. Aspect sentiment classification with both word-level and clause-level attention networks. IJCAI. 2018;2018:4439–45.
31. Majumder N, Bhardwaj R, Poria S, Zadeh A, Gelbukh A, Hussain A, Morency L-P. Improving aspect-level sentiment analysis with aspect extraction. arXiv preprint 2020. arXiv:2005.06607.
32. Dashtipour K, Gogate M, Li J, Jiang F, Kong B, Hussain A. A hybrid persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks. Neurocomputing. 2020;380:1–10.
33. Ma Y, Peng H, Khan TM, Cambria E, Hussain A. Sentic lstm: a hybrid network for targeted aspect-based sentiment analysis. Cogn Comput. 2018;10(4):639–50.
34. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018. arXiv:1810.04805.
35. Ramadhan W, Novianty SA, Setianingsih SC. Sentiment analysis using multinomial logistic regression. In 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC). IEEE 2017. pp. 46–49.
36. Chen T, He T. Benesty m. xgboost: Extreme gradient boosting. R package version 0.4-4. 2016.
37. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint 2013. arXiv:1301.3781.