Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8869385, 19 pages
https://doi.org/10.1155/2020/8869385

Research Article

Big Data Aspect-Based Opinion Mining Using the SLDA and HME-LDA Models

Ling Yuan,[1] **JiaLi Bin**,[1] **YinZhen Wei**,[2] **Fei Huang**,[3] **XiaoFei Hu**,[3] **and Min Tan**[3]

[1]*School of Computer Science, Huazhong University of Science and Technology, 430074, China*
[2]*Huanggang Normal University, 438000, China*
[3]*Wuhan Fiberhome Technical Services Co., Ltd, 430205, China*

Correspondence should be addressed to YinZhen Wei; wyz_gs@163.com

Received 20 July 2020; Revised 1 September 2020; Accepted 23 October 2020; Published 19 November 2020

Academic Editor: Amr Tolba

Copyright © 2020 Ling Yuan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to make better use of massive network comment data for decision-making support of customers and merchants in the big data era, this paper proposes two unsupervised optimized LDA (Latent Dirichlet Allocation) models, namely, SLDA (SentiWordNet WordNet-Latent Dirichlet Allocation) and HME-LDA (Hierarchical Clustering MaxEnt-Latent Dirichlet Allocation), for aspect-based opinion mining. One scheme of each of two optimized models, which both use seed words as topic words and construct the inverted index, is designed to enhance the readability of experiment results. Meanwhile, based on the LDA topic model, we introduce new indicator variables to refine the classification of topics and try to classify the opinion target words and the sentiment opinion words by two different schemes. For better classification effect, the similarity between words and seed words is calculated in two ways to offset the fixed parameters in the standard LDA. In addition, based on the SemEval2016ABSA data set and the Yelp data set, we design comparative experiments with training sets of different sizes and different seed words, which prove that the SLDA and the HME-LDA have better performance on the accuracy, recall value, and harmonic value with unannotated training sets.

1. Introduction

With the development of the Internet, almost all the things of human living have become digitized. The information in network is in the form of structured data and unstructured data [1]. Big data analysis and mining is aimed at discovering implicit, previously unknown, and potentially useful information and knowledge from big databases that contain high volumes of valuable veracious data collected or generated at a high velocity from a wide variety of data sources [2], which is called "4V" of big data. In fact, the deeper mining of big data is to mine the user demand and other deep information; the text mining that this paper studied is one of the ways to mine valid information from big data of text. Therefore, the study proposes two optimized opinion mining methods for customers and merchants to extract valid information they need from massive textual data that satisfies the "4V" [2].

For example, when purchasing a product, people usually refer to others' comments in the specialized product comment area first. Although comments on different platforms have different forms of display, most of them are text-based. Due to most product comments on Taobao only have positive, neutral, and negative labels, what users can directly refer to is just the number of positive and negative comments. Since different users have different needs for the same product, they need to know which attributes of it perform well and which perform poorly. However, these attributes are not shown in detail in the comment interface. It is impossible and uneconomic for users to read more comments to find the attribute they need, because of the massive quantity and continuous growth of product comments. Thus, in order to assist clients and merchants for better decision-making, conducting opinion mining and sentiment analysis of big data is necessary, which will bring huge profits to some markets.

Data mining and analysis have been used in the tourism industry [3], groundwater potential mapping [4], and so on; it becomes more and more important in modern life. Liu and Zhang [5] divided the text mining and analysis tasks into

three levels of granularity: chapter level, sentence level, and attribute level. The chapter level, whose research unit is a document, usually uses algorithms to show whether the opinions expressed by the author are negative or positive, and it is often used to analyse blogs and news. While the sentence-level sentiment analysis is often used to analyse whether the expression of microblogs, tweets, etc., in social networks is negative or positive. Obviously, it is impossible to perform opinion mining at the chapter level and sentence level when analysing user comments. Please consider the following review:

The location of this restaurant is relatively remote, but the waiter has a good attitude and the dishes taste good.

In fact, this sentence contains three opinions. Therefore, we ought not to determine the sentiment polarity of this comment simply and should get more specific results, such as <location, remote, negative>, <service, good, positive>, <dish, good, positive>, i.e., the aspect-based opinion mining. In 2015 and 2016, SEMEVAL released the research topic of ABSA (Aspect-Based Sentiment Analysis), triggering a study boom among scholars. SEMEVAL divides ABSA into three subtasks: identifying entities and attributes, identifying the expression of opinions, and identifying the sentiment polarity of opinions. SEMEVAL also gives the annotated training sets and test sets, such as catering, hotel, and laptop.

In recent years, people devote a lot of energy to analyse comments on the Internet for obtaining the detailed opinions of users [6, 7]. The aspect-based opinion mining of online comments can be divided into two subtasks, namely, the opinion target extraction and the classification of sentiment polarity of comments. Figure 1 gives the overview of the study method classification of the aspect-based opinion mining.

With the development of supervised models, the deep learning, and the conditional random field, CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks) have been widely used in NLP (Natural Language Processing) [8], which have achieved good results in ABSA research. Some other supervised learning methods [9, 10] such as the BMAM [11] which was proposed in recent time also have achieved good results. In addition, Chen et al. [12] and Araque et al. [13] introduced deep learning into the ABSA and achieved satisfactory results. However, compared with the models proposed in this paper based on LDA, all the above models lack the adaptability in various fields and have high manpower costs for annotation. For example, the effects of these models will be greatly reduced when the aspect category of the comment is transferred from the *food* and *beverage* to the *laptop*. Moreover, supervised models such as BMAM [11] need a lot more manpower than the models proposed in this paper to annotate data due to the small number of annotated training sets given. In addition, the time complexity of SLDA and HME-LDA is better than the popular models above. The time complexity of SLDA and HME-LDA is just related to the number of documents, topics, and words, while the time complexity of the popular models such as RNN is related to the multilayer neural network structure, especially the fully connected layer, and the result of the previous time sequence, leading to a really lower speed than the SLDA and HME-LDA.

As for the unsupervised model, there are two basic models for latent semantic analysis: the probabilistic latent semantic analysis (PLSA) [14] model and the latent Dirichlet allocation (LDA) [15] model, which can be applied to extract attributes, assuming that each comment is a combination of attribute words and opinion. Mei et al. [16] proposed a topic-sentiment hybrid model based on the PLSA to extract aspect opinion target words and sentiment words from a group of blogs. Li et al. [17] proposed two kinds of joint models, sentiment LDA and dependence-sentiment LDA, to find positive or negative aspects of sentiment words. Due to the flexibility of the LDA topic model, it is extended and combined with other methods to obtain a topic model [18, 19], which can improve the result topics or the additional information of the model.

In view of the short content, wide coverage and the small number of the annotated corpus of the network comment and its need for aspect-based mining, this paper proposes two schemes based on the LDA topic model that have unsupervised features and good extensibility, making it possible for network comments to perform aspect-based opinion mining with as little annotated data as possible. As for the data cleaning, the amount of data, correctness, completeness, and time correlation [20] are all good evaluation indicators of data quality. As for the data amendment, the Markov Random Field performs well [21].

This paper regards opinion targets, aspects, and opinion expressions as aspect opinion targets that refer to entities or properties to which sentiment words modifies, and sentiment words related to aspect opinion targets are called sentiment opinion words or opinion words.

Moreover, the aspect-based opinion mining schemes proposed in this paper only require users to set corresponding seed words and introduce the classification layer to classify the opinion target words and sentiment opinion words. Meanwhile, in order to improve the effects of the models in schemes, we are biasing the parameters of the LDA model by calculating the similarity between the words in the corpus and the seed words we set.

To sum up, the main study contents of this paper are as follows:

(1) Introduce the NLP tools, WordNet and SentiWordNet, into the standard LDA model to design an optimized LDA-based topic model

(2) Introduce a maximum entropy classifier into the standard LDA model to design another optimized LDA-based topic model

(3) Implement the above two optimized models and design experiments to verify the feasibility and superiority of optimized models

The rest of the paper is organized as follows. Next, we will describe two schemes with two optimized LDA-based models in detail. Then, the experimental results will be given in the next section. At the same time, we will give some analysis about the results. Finally, we give the conclusion.

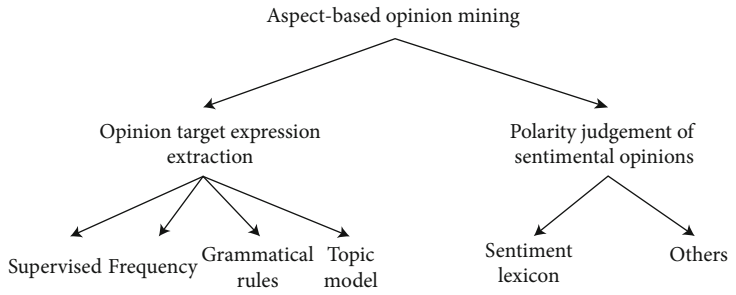

Figure 1: The classification of aspect-based opinion mining.

2. Two Optimized LDA Models for Aspect-Based Opinion Mining

This paper proposes two schemes for aspect-based opinion mining. The first scheme is based on the inverted list and the SLDA (SentiWordNet WordNet-Latent Dirichlet Allocation) model proposed in this paper. The second scheme is based on the inverted list and the HME-LDA (Hierarchical Clustering MaxEnt-Latent Dirichlet Allocation) model proposed in this paper.

The SLDA model is an optimized LDA model based on the WordNet and SentiWordNet, where the WordNet is for the similarity calculation of words and seed words and the SentiWordNet is for the separation of the opinion target words and the opinion words, while the HME-LDA model is an optimized LDA model based on SLDA and MaxEnt-LDA [22]. In fact, SLDA still has the disadvantage of relying on dictionary tools (WordNet and SentiWordNet), leading to the application failure of SLDA in other languages. Luckily, the HME-LDA model solves the problem.

Next, we will illustrate the two optimized models of schemes in detail.

2.1. Optimized Model, SLDA, Based on SentiWordNet and WordNet. Different from the text messages such as documents, blogs, and news, network comments tend to be shorter and often appear in the form of sentences. Please consider the following restaurant comments:

(1) "We, there were four of us, arrived at noon - the place was empty and the staff acted like we were imposing on them and they were very rude."

(2) "Everything is always cooked to perfection, the service is excellent, the decor cool and understated."

In nonaspect opinion mining, the only thing you need to do is to analyse the sentiment polarity of sentences. For example, the word "rude" in the first sentence is negative, so the sentiment polarity of the first sentence is negative. In the second one, the sentiment polarity of the word "perfection" is positive, so the sentiment polarity of the second sentence is positive. This way that only judges the sentiment polarity of sentences does not apply to network comments. More meaningful information should be specific to the word pairs of <Aspect Opinion Target-Opinion> such as <staff-rude>, <cook-perfect>, and <service-excellent>, while the algorithm based on the topic model lacks readability and appointed key words, where the relevant original content fails to be directly found by the final result.

The SLDA is an optimized LDA model based on SentiWordNet (a sentiment dictionary based on WordNet) and WordNet (a large database of English words). Since it is impossible for the LDA itself to separate opinions from opinion targets, this chapter adds a classification layer of opinion words and opinion target words based on the LDA to realize the separation of opinions and aspect opinion targets. The similarity between the word and the seed word in the text, which is reflected on LDA parameters, is calculated by setting the seed word and using the calculation tools of the vocabulary similarity in WordNet. Meanwhile, the opinion target is separated from the sentiment opinion words using tools that calculate the vocabulary sentiment in SentiWordNet. Aiming at the lack of readability of LDA results, we establish a belonging relationship among the clustering results, seed words, and original texts. Also, the SLDA model needs to set seed words, but has no need for the additional annotated data sets.

In order to achieve the goal that users can quickly find what they want from massive comments by inquiring the index rather than by reading, there are three steps to do:

(1) Construct an inverted index to number sentences and words

(2) Determine the aspect category of comments, separate the aspect-based attribute words from the sentiment opinion words, and determine the aspect category to which they belong

(3) Enhance the readability of results. Users can see an overview of Domain-Aspect-Opinion and find the specific sentence by the inverted index of a word

It is worth noting that the premise of our study is that the aspect category of the original corpus is known. The aspect category, which has the belonging relationship with the aspect opinion target, is usually given in advance by the original corpus, which can be learned by the user guide of the corpus. For example, the aspect opinion target "steak" belongs to the aspect category "Food." In addition, both words and sentences can belong to an aspect category. Thus, what only need to do is to label aspect opinion target words and sentences with a specific aspect category.

2.1.1. Implementation Process. Figure 2 is an overall flow chart of the scheme that conducts aspect-based opinion mining based on SLDA and the inverted index.

(1) Construct an inverted index. The words in the corpus are numbered in the form of binary group $< a, b >$, where a is the serial number of the sentence and b is the serial number of the words in the sentence. In addition, the generation of the inverted list requires the removal of duplicate word and the recording of their numbers. The inverted list reserves the sentence number that contains the word and the position information of the word in the sentence, making it easier to retrieve with context information from the original corpus later

(2) Data preprocessing, whose main tasks are to extract the data required and remove stop words from the original corpus for making a training set. The formats of the original data sets are XML and CSV. It is necessary to extract comment statements by the corresponding labels and fields. And the text in the corpus contains many useless stop words, such as "is" and "a," which should be removed before further processing to avoid interference with the training of the SLDA model

(3) Introduce preprocessed data into SLDA for model training and get clustering results. The setup of seed words, as well as the assist of WordNet and SentiWordNet, is required in this process

(4) Process the clustering results for better readability. The results in the topic model are probability matrix, whose readability is poor. To solve this problem, it is necessary to find the word corresponding to the result with higher probability and find its original sentence by the inverted index

In short, the SLDA model, trained with the preprocessed data of step 2 above, queries the original sentences that contain keywords from the original corpus by the inverted index of step 1 above.

2.1.2. The Optimized Direction of LDA. The expectation of every random variable μ_i of the Dirichlet distribution can be expressed as $E(\mu_i)$. The value of $E(\mu_i)$ can be calculated by equation (1), where α represents the parameter of the Dirichlet distribution and K represents the number of topics:

$$E(\mu_i) = \frac{\alpha_i}{\sum_{i=1}^{K} \alpha_i}, \quad (1)$$

where α is a fixed value and the $E(\mu_i)$ of each topic is same. Based on this, a biasing method of α can be explored to make the expectations of the corresponding probabilities of each topic different, so as to generate the topic bias. More visually, when α is fixed, it means that we fail to know which topic word to use before generating the document. When α is biased, the ideal topic is determined before the document is generated.

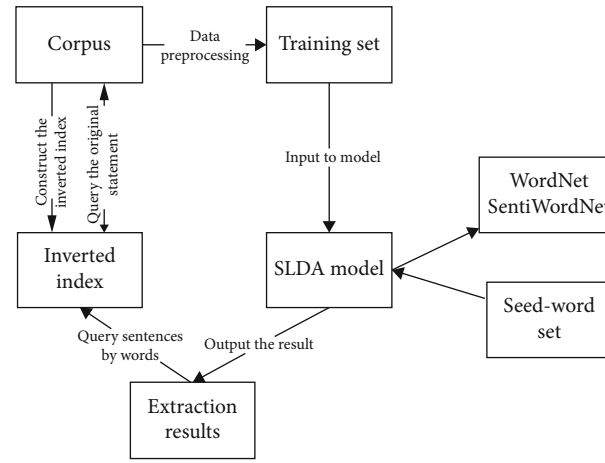

Figure 2: The overall flow chart of the scheme that is based on SLDA and inverted index.

Actually, the topic number $Z_{m,n}$ can be used as an indicator variable in the LDA model to control the selection of topic-word distribution. Based on this, more indicator variables can be introduced to refine the topic and obtain more topic-word distributions.

From the above, there are two aspect opinion targets that the LDA model can optimize: the first is to bias the parameters, α and β, which can generate topic biases to improve the classification effect; the second is to introduce more indicator variables like $Z_{m,n}$, which can generate more detailed topic classifications.

2.1.3. The Description of the SLDA Model. The standard LDA uses the document as the unit of topic allocation, while in the aspect-based opinion mining of the network comments, the sentence is often used as the unit of topic allocation, because there is no document with large contents in network comments and the topic allocation of vocabulary in network comments is actually more meaningful. In the aspect-based opinion mining of network comments, it is necessary to extract the aspect category of the comment, the opinion target, and the comment opinion (sentiment polarity) from the text. For example, in restaurant comments, "food," "service," and "ambiance" are the aspect categories of comments. In the "food" category, "steak" is the opinion target, and the evaluation of "good" for "steak" is the opinion of the comments (sentiment polarity).

In the SLDA model, seed words are directly used as aspect categories of comments, while the opinion targets of comments and the comment opinions are separated by the introduced sentiment layer, and the positive and negative polarities of comments are classified as well. The PGM (Probabilistic Graphical Model) of SLDA is shown in Figure 3.

2.1.4. The Generation Process of the SLDA Model. Only the parameters α and β, which, respectively, belong to the document-topic Dirichlet distribution and the topic-word

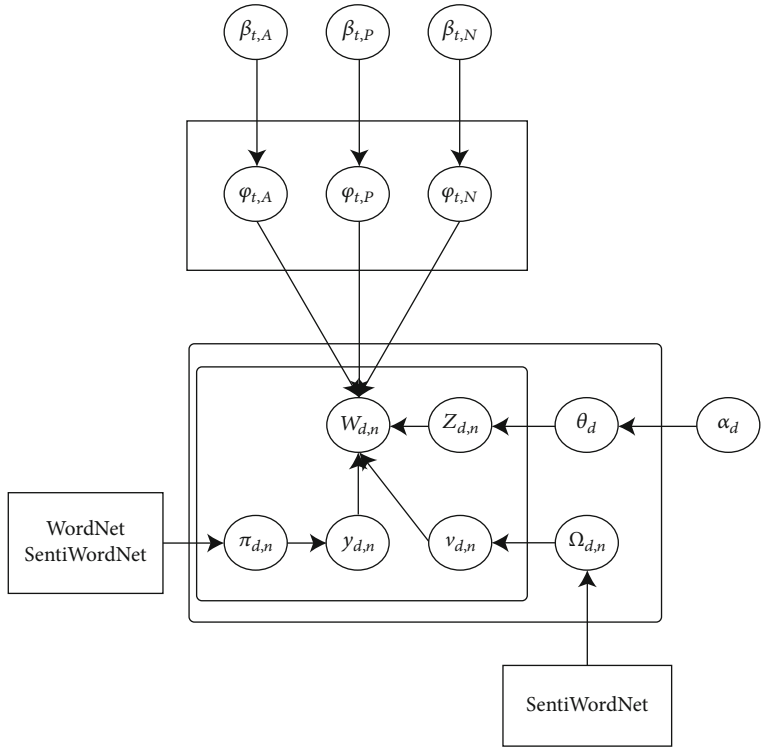

Figure 3: The probabilistic graphical model of SLDA.

Dirichlet distribution, are introduced into the standard LDA. In the SLDA model, seed words have been identified as the aspect category of comments. The variable $y_d \in \{A, O\}$ is introduced to represent the separation of the opinion target and the comment opinion. When $y_d = A$, the current word is the opinion target of comments. When $y_d = O$, the current word is the comment opinion. Meanwhile, the variable $v_d \in \{P, N\}$ is introduced into SLDA. When $v_d = P$, the sentiment polarity of the current vocabulary is positive, and when $v_d = N$, the sentiment polarity of the current vocabulary is negative. Both y_d and v_d are determined by the corresponding algorithms based on the WordNet and SentiWordNet. In the standard LDA model, the values of parameters, α and β, are fixed. In the SLDA model, different α_d will be set for each sentence, and the parameters, $\beta_{t,A}$, $\beta_{t,P}$, and $\beta_{t,N}$, will be set for the opinion target, the positive comments and the negative comments, respectively.

In SLDA, the first step is to generate the sentence-topic distribution and determine a topic for each sentence by the multinomial distribution. Then, two influence factors, y_d and v_d, are determined by WordNet and SentiWordNet to indicate the aspect categories of words. By the SLDA generation method above, we finally select a topic-word distribution and determine the final word.

The concrete generation process of SLDA is as follows:

Firstly, the distributions of the opinion target $\varphi_{t,A}$, the positive opinion word $\varphi_{t,P}$, and the negative opinion word $\varphi_{t,N}$ are, respectively, extracted from the parameters, $\beta_{t,A}$, $\beta_{t,P}$, $\beta_{t,N}$, where $\varphi_{t,A} \sim \mathrm{Dir}(\beta_{t,A})$, $\varphi_{t,P} \sim \mathrm{Dir}(\beta_{t,P})$, and $\varphi_{t,N} \sim \mathrm{Dir}(\beta_{t,N})$.

Similar to the standard LDA, for each sentence (document in the standard LDA), a topic distribution, $\theta_d \sim \mathrm{Dir}(\alpha_d)$, is extracted from the Dirichlet distribution with the parameter α_d.

Then, the word $w_{d,n}$ in the sentence d extracts a topic number t, i.e., extracts $z_{d,n} \sim \mathrm{Multi}(\theta_d)$.

After determining the topic number of the word $w(d, n)$, the classification of it remains to be determined. After determining the topic t in SLDA, in order to further classify the word as the opinion target word or the opinion of comments, the variables, $y_{d,n} \in \{A, O\}$ and $v_{d,n} \in P, N$, are introduced into the SLDA model. The $y_{d,n}$ and $v_{d,n}$ point to the n-th word $w_{d,n}$ in the sentence d jointly. The $y_{d,n}$ is used to indicate that the word $w_{d,n}$ is the opinion target of comments or the comment opinion, which is extracted from the Bernoulli distribution on $\{0, 1\}$ with the parameter $\pi_{d,n}$. The $v_{d,n}$ is used to indicate that the word $w_{d,n}$ is a positive or a negative comment opinion, which is extracted from the Bernoulli distribution on $\{0, 1\}$ with the parameter $\Omega_{d,n}$. The above two Bernoulli distributions are calculated by the WordNet and SentiWordNet. Finally, the word $w_{d,n}$ can be extracted according to equation (2):

$$Wd, n \sim \begin{cases} \mathrm{Multi}(\varphi t, A), & \text{if } yd, n = A, \\ \mathrm{Multi}(\varphi t, A), & \text{if } yd, n = O, \text{and } vd, n = P, \\ \mathrm{Multi}(\varphi t, A), & \text{if } yd, n = O, \text{and } vd, n = N. \end{cases} \tag{2}$$

Table 1 gives the description of the related symbols in the SLDA model, which is useful for readers in reading.

2.1.5. The Inference Process of the SLDA Model. The SLDA model consists of two major parts. One is the classification of the opinion target words and sentiment opinion words composed of the WordNet and SentiWordNet as well as the classification of positive and negative sentiment opinion words. The other is the LDA topic model. Besides, the seed words should be set before the inference of SLDA. Next, several modules will be introduced in turn.

(1) The Setting of Seed Words. In SLDA, seed words are set as aspect categories of comments. For example, in the classic English comment set of the Restaurant, the aspect categories are *food*, *service*, *ambiance*, etc. If the corpus is the Restaurant English comment set, the seed words can be directly set as *food*, *service*, and *ambience*. The seed word is recorded as w_t, where $t \in \{1, \cdots, T\}$, that is to say, the number of seed words determines the number of topics in the SLDA model.

(2) The Inference of the Word Classification Model Based on the WordNet and SentiWordNet. In the SLDA model, the Bernoulli distribution with parameter $\pi_{d,n}$ and Bernoulli distribution with parameter $\Omega_{d,n}$, which are, respectively, used to separate the opinion target words from sentiment opinion words and separate positive sentiment opinion words from negative sentiment opinion words, are related to the WordNet and SentiWordNet. The calculation of $\pi_{d,n}$ depends on the seed word w_t.

The words in WordNet have the feature of polysemy. To calculate the similarity between words, it is necessary to determine the exact meaning of a word. Therefore, before the model inference, it is necessary to determine the semantic interpretation $s_{t,k0}$ of the seed word w_t in the WordNet. When the current word is $w_{d,n}$, its semantic interpretation in the WordNet is $s_{d,n,k}$, where $k \in \{1, \cdots, K\}$ and K is the number of semantic interpretations. After determining the semantics of the seed words, we can regard the seed word as the topic and the aspect category of the final opinion target. All nonsentiment opinion words will be grouped into a certain aspect category of comments. Therefore, the similarity between the semantics of the current word and the semantics of the seed word can be calculated, and the semantics with the greatest similarity can be determined as the meaning expressed by the word in the sentence finally. The degree of semantic similarity between different words in the WordNet is recorded as $\mathrm{Sim}(s_1, s_2)$; then, the semantic similarity, $\mathrm{Sim}(s_{d,n,k}, s_{t,k0})$, between the $s_{d,n,k}$ of $w_{d,n}$ and the $s_{t,k0}$ of each seed word w_t can be calculated, where $k \in \{1, \cdots, K\}$. Besides, K is the number of semantic interpretations, $t \in \{1, \cdots, T\}$, and T is the number of seed words. In all calculation results, we choose the semantics with the max value of the similarity result and determine the largest result k' as the semantics $s_{d,n,k'}$ to which the current word $w_{d,n}$ belongs.

After determining the semantic $s_{d,n,k'}$ of the word $w_{d,n}$, the sentiment polarity of $s_{d,n,k'}$ can be queried in the SentiWordNet. In the SentiWordNet, the semantics of a word has three sentiment propensity probabilities: ρ_o indicates the probability that the semantics is objective (excluding sentiment polarity), p_P indicates the probability that the semantics is positive, and ρ_N indicates the probability that the semantics is negative. Besides, $p_o + p_P + p_N = 1$. It is believed that if the semantics is objective, the word is the opinion target, whose corresponding probability is p_o; otherwise, it is a sentiment opinion word. The sentiment scores of the semantics $s_{d,n,k'}$ of the word $w_{d,n}$ are recorded as $p^o_{d,n,k}$, $p^P_{d,n,k}$, and $p^N_{d,n,k}$. In Section 2.1.4, the Bernoulli distributions with parameter $\pi_{d,n}$ and parameter $\Omega_{d,n}$ can be determined by equation (3) and equation (4):

$$\pi_{d,n} = p^o_{d,n,k}, \tag{3}$$

$$\Omega_{d,n} = \frac{p^P_{d,n,k}}{p^P_{d,n,k} + p^N_{d,n,k}}. \tag{4}$$

So far, the model inference based on the WordNet and SentiWordNet has been completed. Algorithm 1 is the pseudocode for calculating $\pi_{d,n}$, $\Omega_{d,n}$.

(3) The Inference of the SLDA Model. In the standard LDA, the parameters, α and β, of the Dirichlet distribution are fixed. While in the SLDA model, the parameters, α and β, are biased by calculating the similarity between the input corpus and seed words. The fixed parameters are recorded as α_{base} and β_{base}, and the biased parameters are recorded as α_d, $\beta_{t,A}$, $\beta_{t,P}$, and $\beta_{t,N}$. In this paragraph, the semantic similarity between the word w and the seed word t is recorded as $\mathrm{sim}(w, t)$. The probability that w is a positive word is recorded as $\mathrm{sim}(w, P)$. The probability that w is a negative word is recorded as $\mathrm{sim}(w, N)$. The w is recorded as $\mathrm{sim}(w, A)$ if it belongs to the opinion target word. The w is recorded as $\mathrm{sim}(w, O)$ if it belongs to the sentiment opinion word.

In the standard LDA, the parameter α is used to control the topic distribution probability of the document. For all documents in the corpus, the values of α are the same, leading to determine which topic word will be the document topic before generating the document hardly. However, the parameter α_d in SLDA, which can be calculated by equation (5), will be set separately for each document (for each sentence in SLDA) based on the similarity between vocabulary and seed words, leading to determine a more ideal topic in advance before generating the document.

$$\alpha_d = \frac{\sum_i^{N_d} \mathrm{sim}(w_{d,i}, t)}{\sum_{t'}^{T} \sum_i^{N_d} \mathrm{sim}\left(w_{d,i}, t'\right)} \times \alpha_{\text{base}}. \tag{5}$$

In equation (5), N_d is the number of all words in the current sentence, T is the number of topics, $w_{d,i}$ is the i-th word in the current sentence, and t is the seed word.

In the standard LDA, the parameter β is used to control the word distribution of each topic, and the value of β is the same for each topic. In SLDA, parameters which can be

Table 1: The description of related symbols in the SLDA model.

Symbol	Description
D	The total number of comments in the corpus. The unit of corpus is the sentence
T	The number of topics
V	The number of words in the corpus
$w_{d,n}$	The n-th word in the d-th comment in the corpus
$z_{d,n}$	The topic of d-th comment. The value is $\{1, \cdots, T\}$
$y_{d,n}$	An indicator variable. The value is $\{A, O\}$. It is used to indicate the opinion target words and the sentiment opinion words.
$v_{d,n}$	An indicator variable. The value is $\{P, N\}$. It is used to indicate positive and negative sentiment opinion words.
A, O, P, N	The opinion target, the sentiment opinion word, the positive sentiment word, the negative sentiment word
$\varphi_{t,A}$	The distribution of the opinion target generated by a priori Dirichlet distribution with the parameter $\beta_{t,A}$
$\varphi_{t,P}$	The distribution of positive comment words generated by a priori Dirichlet distribution with parameters $\beta_{t,P}$
$\varphi_{t,N}$	The distribution of negative comment words generated by a priori Dirichlet distribution with parameters $\beta_{t,N}$
θ_d	The distribution of sentence topic terms generated by a priori Dirichlet distribution with parameter α_d

1: Query the semantic list S_{list} of $w_{d,n}$ in WordNet
2: //*Record the semantic value of the max similarity*
3: $s_w = 0$;
4: //*Record the maximum similarity*
5: $sim_{\max} = 0$;
6: //*Each semantic s of* $w_{d,n}$
7: **for** $s \in s_{list}$ **do**
8: **for** $s_t \in s_{t_{list}}$ **do**
9: //*Sim() is a function provided by WordNet*
10: **if** $\text{sim}_{\max} < \text{Sim}(s, s_t)$ **then**
11: $s_w = s$;
12: $sim\_\max = Sim(s, s_t)$;
13; **end if**
14: **end for**
15: **end for**
16: Use SentiWordNet to query the sentiment polarity of semantic S_w
17: Calculate $\pi_{d,n}$ using equation (3)
18: Calculate $\Omega_{d,n}$ using equation (4)
19: **return** s_w
20: **return** $\pi_{d,n}$
21: **return** $\Omega_{d,n}$

Algorithm 1. The calculation of $\pi_{d,n}$, $\Omega_{d,n}$.

calculated by equation (6), equation (7), and equation (8) will be set separately for each topic based on the similarity between vocabulary and seed words.

$$\beta_{t,A} = \text{sim}(w, A) \times \beta_{\text{base}}, \tag{6}$$

$$\beta_{t,P} = \text{sim}(w, P) \times \beta_{\text{base}}, \tag{7}$$

$$\beta_{t,N} = \text{sim}(w, N) \times \beta_{\text{base}}. \tag{8}$$

Similar to the standard LDA, SLDA is solved by the Gibbs sampling method. The variables involved in the solution process are explained in Table 2.

Equation (9) is used to sample the topic of each sentence.

$$p\left(Z_{d,n} = t \middle| z_{\neg d,n}, y_{\neg d,n}, v_{\neg d,n}, \cdot\right) \propto \frac{n_{w_{d,n}}^{t,A} + \beta_{w_{d,n}}^{t,A}}{\sum_v^V \left(n_v^{t,A} + \beta_v^{t,A}\right)} \times \frac{n_{w_{d,n}}^{t,P} + \beta_{w_{d,n}}^{t,P}}{\sum_v^V \left(n_v^{t,P} + \beta_v^{t,P}\right)} \times \frac{n_{w_{d,n}}^{t,N} + \beta_{w_{d,n}}^{t,N}}{\sum_v^V \left(n_v^{t,N} + \beta_v^{t,N}\right)} \times \left(n_{d,t} + \alpha_{d,t}\right). \tag{9}$$

Table 2: The partial symbolic description of SLDA model inference.

Symbol	Description
$n_v^{t,A}$	The number of words v in topic t and category A
$n_{d,t}$	The number of words with the topic t in the d-th sentence
$\beta_v^{t,u}$	The number of words v in topic t and category A

Equation (10) and equation (11) are used to sample $y_{d,n}$ and $v_{d,n}$.

$$p\left(y_{d,n}=u\middle|z_{d,n}=t,\cdot\right)\alpha\frac{\left(n_{w_{d,n}}^{t,u}+\beta_{w_{d,n}}^{t,u}\right)\times\mathrm{sim}(w_{d,n},u)}{\sum_v^V\left(n_v^{t,u}+\beta_v^{t,u}\right)},\ u\in\{A,O\},\tag{10}$$

$$p\left(v_{d,n}=q\middle|z_{d,n}=t,\cdot\right)\propto\frac{\left(n_{w_{d,n}}^{t,q}+\beta_{w_{d,n}}^{t,q}\right)\times\mathrm{sim}(w_{d,n},q)}{\sum_v^V\left(n_v^{t,q}+\beta_v^{t,q}\right)},\ q\in\{P,N\}.\tag{11}$$

In the corpus, the approximate probability of the topic t and the sentence d can be calculated by equation (12).

$$\theta_d=\frac{n_{d,t}+\alpha_{d,t}}{n_d+\sum_{t'}^T\alpha_{d,t'}}.\tag{12}$$

With t as the topic, the approximate probability that the word $w_{d,n}$ is the opinion target can be calculated by equation (13), the approximate probability that the word $w_{d,n}$ is the positive opinion can be calculated by equation (14), and the approximate probability that the word $w_{d,n}$ is the negative opinion can be calculated by equation (15).

$$\varphi_{w_{d,n}}^{t,A}=\frac{n_{w_{d,n}}^{t,A}+\beta_{w_{d,n}}^{t,A}}{\sum_v^V\left(n_v^{t,A}+\beta_v^{t,A}\right)},\tag{13}$$

$$\varphi_{w_{d,n}}^{t,P}=\frac{n_{w_{d,n}}^{t,P}+\beta_{w_{d,n}}^{t,P}}{\sum_v^V\left(n_v^{t,P}+\beta_v^{t,P}\right)},\tag{14}$$

$$\varphi_{w_{d,n}}^{t,N}=\frac{n_{w_{d,n}}^{t,N}+\beta_{w_{d,n}}^{t,N}}{\sum_v^V\left(n_v^{t,N}+\beta_v^{t,N}\right)}.\tag{15}$$

(4) Gibbs Sampling Implementation of the SLDA Model. The Gibbs sampling process of SLDA mainly has the following steps:

(1) *Random initialization*: randomly assign a topic number z to all sentences in the corpus, the topic numbers of all words in the sentence are also set to z, then the values of the indicator variables, y and u, are randomly set for all words in the sentence, where $y\in\{A,O\}$, $u\in\{P,N\}$.

(2) Traverse the corpus again, resample the topics of all words according to equation (9), and update the relevant values. Then, resample the indicator variables, y and u, according to equation (10) and equation (11), and update their values

(3) Repeat step 2 until the Gibbs sampling results converge

(4) Process the final results to improve the readability and save the results

In the initialization phase of the Gibbs sampling, a topic number is randomly assigned to each sentence in the corpus. For each word in the sentence, three categories are randomly assigned, which are determined by the indicator variables, y and u. Then, add 1 to the corresponding statistics. A part of the pseudocode in the initialization phase is shown in Algorithm 2.

In the repeated iteration phase of Gibbs sampling, the topic of each sentence in the corpus and the category of each word in the sentence are resampled. Then, update the relevant statistics after each sampling. Repeat the process until the end of the iteration.

In the result processing stage, the relevant value φ can be calculated by equation (13), equation (14), and equation (15). The calculated φ value is a digital, and the variables required in the calculation are the topic number, the category to which the word belongs, and the number of the word in the vocabulary. The relevant sentence information is missing; accordingly, it is necessary to effectively organize various types of information for the user to view. We use *result* to represent the final result. The *result.topic* is the topic information, that is, the seed word set by ourselves, which also can be regarded as the comment category word. The *result.word* saves the original word. The *result.wordType* is the category of the current word (aspect target words, positive and negative opinion words). The *result.sentences* is the sentence to which the word belongs. The *result.prob* is the probability that the word becomes a member of the category which the comment belongs to. The first m results are generated for each category under all the topics, and the relevant pseudocode is shown in Algorithm 3.

After getting the finalResults, we can query the results according to both the topic and the wordType.

2.2. Optimized Model, HME-LDA, Based on MAXENT-LDA. Because the WordNet and SentiWordNet only support English, SLDA has no linguistic adaptation. Therefore, we propose an optimized model in this chapter, namely, the HME-LDA model that has the linguistic adaptation. The

```
1: for d = 1 to D do
2:     //Randomly assign topic t to sentences in Corpus
3:     t = randomint from (1, T);
4:     documentTopics [d] = t;
5:     documentTopicsCount [d][t]++;
6:     for w = 1 to N do
7:       //Randomly assign value to y
8:       y = randomint from (0,1);
9:       Y[d][w] = y;
10:      //Randomly assign value to u
11:      u = randomint from (0,1);
12:      U[d][w] = u;
13:      if y ==0 then
14:        //statistic of aspect targets with topic t plus 1
15:        aspectWordCount [w][t]++;
16:      end if
17:      if y ==1 and u ==0 then
18:        //statistic of positive opinion with topic t plus 1
19:        positiveWordCount [w][t]++;
20:      end if
21:      if y ==1 and u ==1 then
22:        //statistic of negative opinion with topic t plus 1
23:        negativeWordCount [w][t]++;
24:      end if
25:    end for
26: end for
```

Algorithm 2. The initialization of Gibbs sampling.

optimized model, HME-LDA, is proposed by combining with the MaxEnt-LDA [22] model and the SLDA model. And the HME-LDA uses an unsupervised hierarchical clustering method to generate the annotated data set required by the maximum entropy model. Based on these, the new model can be used for comment opinion mining in many other languages.

2.2.1. Implementation Process. The overall process of the scheme that conducts aspect-based opinion mining based on the HME-LDA and the inverted index is shown in Figure 4:

(1) This step is the same as the implementation of step 1 in Section 2.1.1

(2) Data preprocessing, whose main task is to remove stop words. The text in the corpus contains many useless stop words, such as "is" and "a," which should be removed before further processing to avoid interference with the results

(3) Automatically generate annotated data sets by hierarchical clustering and train maximum entropy models for classification of the opinion target words and sentiment opinion words

(4) Enter the data into the HME-LDA model and perform the training to get the results

(5) Process the results for better readability. The results in the topic model are probability matrixes, whose readability is poor. To solve this problem, it is necessary to find the word which corresponds to the result that has a higher probability and find its original sentence by the inverted index

As for Step 3 above, there are a lot of word order information in the corpus; thus, when the category of a word in the corpus can be determined, the feature $\{w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}\}$ can be extracted. In HME-LDA, seed words are divided into topic seed words of aspect categories, seed words expressing positive sentiment, and seed words expressing negative sentiment. By scanning corpus, annotated feature sets, $\{w_{i-2}, w_{i-1}, w_i^u, w_{i+1}, w_{i+2}\}$, can be obtained, where $u \in \{A, O\}$. When $u = A$, w_i^A is the topic seed words, and when $u = O$, w_i^O is the sentiment seed words, and u is regarded as the label. However, the number of seed words is limited, so the training set obtained by scanning corpus may be insufficient in size. Therefore, the words in the corpus are considered to be clustered, and all the words in the category of the seed words are considered to have the same category as the seed words, and the word w_i^u in the scanned annotated feature set $\{w_{i-2}, w_{i-1}, w_i^u, w_{i+1}, w_{i+2}\}$ is replaced with the word in the category of seed words to get the new annotated data.

2.2.2. MaxEnt-LDA Model. In order to realize the aspect-based opinion mining of reviews, Zhao et al. [22] proposed a MaxEnt-LDA model based on LDA. Figure 5 is the probability model diagram of MaxEnt-LDA [22].

Adopting the second thought of optimizing LDA in Section 2.1.2, MaxEnt-LDA further divides topics and increases the topic number by introducing another two indicator variables, $y_{d,s,n}$ and $u_{d,s,n}$, that are similar to $Z_{m,n}$. Three new topic-word distributions are generated by the Dirichlet distribution with the parameter β. Meanwhile, the original topic is further divided into categories A (aspect-term) and O (opinion word).

In the MaxEnt-LDA model, the indicator variable $y_{d,s,n}$ is generated by the sampling of the maximum entropy model. In this model, the selected features of the maximum entropy model include the word and the part of speech, while there are three labels which are background word B, opinion word O, and the opinion target word A. The annotated training set is partially extracted from the SemEval data set and partially annotated by manual. The indicator variable $y_{d,s,n}$ is the same as $Z_{m,n}$, both of which are sampled from the multinomial distribution generated by the Dirichlet distribution.

MaxEnt-LDA increases both the type and number of classification and introduces the maximum entropy model. However, the MaxEnt-LDA model increases the dependence on the annotated data simultaneously. In order to pursue the unsupervised features of the model, there are two ways to improve the MaxEnt-LDA: one is to replace the maximum entropy model with other unsupervised classification models; the other is to use unsupervised methods to automatically label data sets to avoid the dependence on annotated data. Meanwhile, the parameters α and β can be considered for bias.

2.2.3. The Description of the HME-LDA Model. In the HME-LDA model, seed words are directly set as an aspect category

```
1: //Create a 3d array to store results
2: init results[][][]
3: for topic in TopicList do
4:   for wordType in {A, P, N} do
5:     for v in V do
6:       result = new result
7:       result.topic = topic
8:       result.word = v
9:       result.wordType = wordType
10:        Calculate φ according to Equation. (13), Equation. (14), Equation. (15)
11:        result.prob = φ
12:        //Add result to the array
13:        results[topic][wordType][v] = result
14:    end for
15:  end for
16: end for
17: //Next, sort the results
18: for topic in TopicList do
19:  /Store sentences contain results
20:  finalResults[][][]
21:  for wordType in {A, P, N} do
22:    //Sort by φ value and get the results of the top m
23:    finalResults[topic][wordType] = getTopKByPhi(results[topic][wordType], m)
24:    fork = 1 to m do
25:      Query the corresponding sentence of result.v from the inverted index described in Section 3.1.1 and add it to the result
26:    end for
27:  end for
28: end for
29: return finalResults
```

Algorithm 3. The result processing of the Gibbs sampling.

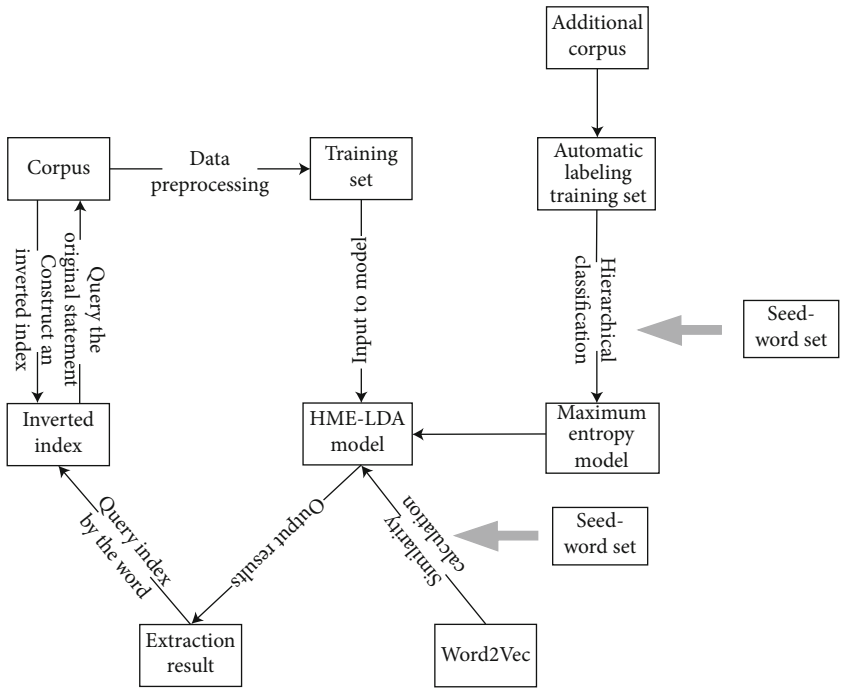

Figure 4: The flow chart of the scheme that is based on the HME-LDA and inverted index.

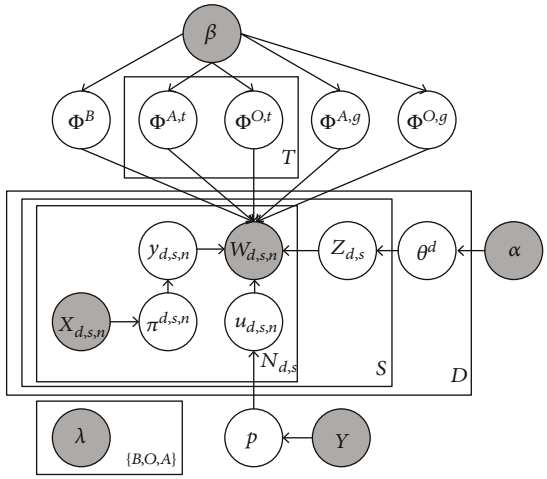

Figure 5: The probability diagram of MaxEnt-LDA.

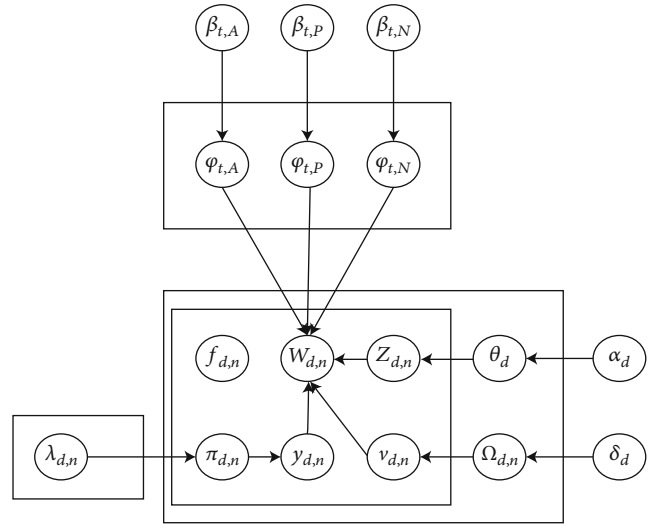

Figure 6: The probabilistic graphical model of HME-LDA.

of reviews. Therefore, it only needs to classify the opinion targets of reviews and the review opinions by introducing the maximum entropy model as a classifier. The Bernoulli distribution with the parameter $\pi_{d,n}$, where d indicates the d-th sentence and n indicates the n-th word, is jointly determined by the weight $\lambda_{d,n}$ and the eigenvector $f_{d,n}$ of the maximum entropy model. A beta distribution with a parameter δ_d will be introduced as a priori to generate a Bernoulli distribution with a parameter Ω_d for the classification of positive and negative sentiment opinion words. Figure 6 is the PGM.

The generation process of the HME-LDA model is similar to that of the SLDA model in Section 2.1. The difference is that $y_{d,n}$ is determined by the maximum entropy model rather than by the WordNet and SentiWordNet. And $v_{d,n}$ is determined by the parameter δ_d rather than by the calculation of the WordNet and SentiWordNet. The generation process of the HME-LDA model can refer to Section 2.1.4.

In addition, because there are no methods to calculate the sentiment polarity in HME-LDA, it is necessary to set a sentiment opinion word whose sentiment polarity is positive or negative for each comment category.

2.2.4. The Automatic Data Annotation Method Based on Hierarchical Clustering. The MaxEnt-LDA [22] model uses the word feature with a window whose size is 3 to extract the features from the annotated words. The selected features include the word and the word order $\{w_{i-1}, w_i, w_{i+1}\}$, where w_i is the current word. The selected features also include the features of grammatical rules of words $\{POS_{i-1}, POS_i, POS_{i+1}\}$, where POS_i indicates the part of speech of the current word (adjectives, nouns, verbs, etc.). The part-of-speech tagging requires the use of additional tools, while it is different for the accuracy of part-of-speech tagging due to different kinds of languages, and there is a possibility that the tagging tool is lacking. Therefore, the features selected in this chapter are only the words themselves.

(1) The Process of Automatically Annotating Data. After identifying the feature information obtained from the training, the next step is to consider how to automatically label it. In order to extract the feature $\{w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}\}$, the only thing that needs to do here is to determine the category of the word w_i in the corpus with information of word order. The seed-word setting of HEM-LDA is explained in Section 2.2.3. In the HME-LDA, seed words can be divided into three categories: topic seed words of a review category, seed words expressing positive sentiments, and seed words expressing negative sentiments. And the opinion target words belong to the corresponding category of comments, both of which are the same kind of topic. Therefore, seed words can be divided into two categories: opinion target words and sentiment opinion words. The categories of these seed words are able to be determined. The annotated feature sets $\{w_{i-2}, w_{i-1}, w_i^u, w_{i+1}, w_{i+2}\}$ can be obtained by scanning the corpus, where $u \in \{A, O\}$. When $u = A$, w_i^A is the topic seed word. When $u = O$, w_i^O is the sentiment seed word, and u is the label.

However, the number of seed words is limited, and the size of the training set obtained by scanning corpus may be insufficient. Therefore, we attempt to cluster the words in the corpus and treat that all words in the category of seed words have the same category as the seed words. Besides, the word in this category is used to replace the word w_i^u in the scanned annotated feature set $\{w_{i-2}, w_{i-1}, w_i^u, w_{i+1}, w_{i+2}\}$, so as to obtain the new annotated data. The pseudocode is shown in Algorithm 4.

(2) The Selection of Clustering Method. In this section, in order to improve the domain adaptability of the model and avoid the different effects of the same value of K when using the data of different fields, and to avoid more parameter adjustments as well, we select the hierarchical clustering method which has no need for the number of clusters. The result of hierarchical clustering is shown in Figure 7.

In the tree structure generated by the hierarchical clustering results, the leaf nodes of the tree are words in the corpus.

```
1: for t in Topics do        //Process each seed word
2:   wordList = getWrodListFromCorpus(t)    //Find the location where t appears from the corpus and get the corresponding word order
3:   wordCluster = getWordCluster(t)    //Get all the words of the category t from the clustering results
4:   trainSet = new Set    //Used to save labelled training samples
5:   for wOrder in wordList do
6:     for w in wordCluster do
7:       replaceWord(wOrder, t, w)    //Replace the word t in the word order with w
8:       trainSet.add(wOrder, t.Type)    //Add the label to which wOrder and t belong to the training set
9:     end for
10:  end for
11: end for
```

Algorithm 4. The process of automatically labelling data.

When looking for words of the same category, the intermediate node of the upper layer can be found by the current word. All the leaf nodes of the subtree with this intermediate node as the root node belong to the same category.

2.2.5. The Inference of the HEM-LDA Model. The reasoning of the HME-LDA model mainly includes two parts. The first part is the inference of the maximum entropy model for the classification of the opinion target words and sentiment opinion words. And the second part is the inference of the optimized LDA model.

(1) The Maximum Entropy Model. The maximum entropy model solves the classification problem actually. When the input of the model is x, the probability $p(y \mid x)$ of the category y can be calculated by equation (16).

$$p(y \mid x) = \frac{e^{\sum_{i=1}^{n} \lambda_i f_i(x,y)}}{\sum_y e^{\sum_{i=1}^{n} \lambda_i f_i(x,y)}}. \tag{16}$$

In equation (16), λ_i represents the weight vector, $f_i(x, y)$ is the eigenfunction, and n is the number of categories. When using the maximum entropy model for the classification of the opinion target words and sentiment opinion words, it is necessary to select appropriate features. The MaxEnt-LDA [22] model uses two features, i.e., word order and part of speech. The part of speech tagging relies on some other tools that is different with different languages. In order to avoid using tools that rely on languages, this chapter chooses the word order as a feature. Section 2.2.4 gives the method of automatically annotating the training set.

By training the maximum entropy model, the weight λ_u of the feature set $f_{d,n}$ can be obtained, and $\pi_{d,n}^{u}$ can be obtained by equation (17), where d represents the d-th sentence, n is the n-th word in the sentence, $u \in \{0, 1\}$ is the label collection, 0 represents the opinion target whose type is A, and 1 represents the sentiment opinion whose type is O.

$$p\left(y_{d,n} \mid f_{d,n}\right) = \pi_{d,n}^{u} = \frac{e^{\lambda_u \times f_{d,n}}}{\sum_{i=0}^{1} e^{\lambda_i \times f_{d,n}}}. \tag{17}$$

(2) The Model Inference of HME-LDA. In SLDA, the parameters, α and β, are offset by calculating the similarity between the input corpus and the seed words. The fixed parameters are denoted as α_{base} and β_{base}, and the offset parameters are denoted as α_d, $\beta_{t,A}$, $\beta_{t,P}$, and $\beta_{t,N}$. In Chapter 2, the above parameters are offset by using the semantic similarity calculation of the WordNet and the sentiment polarity calculation of the SentiWordNet. In this section, both the Word2Vec model and the cosine distance are used to calculate the similarity between words. The training corpus of the Word2Vec model is the all content in the corpus. The vector of the current word in the Word2Vec is represented by v_w, and the vector of the seed word w_t whose topic is t in Word2Vec is represented by v_{w_t}. The similarity between the word w and the topic word w_t can be calculated by equation (18).

$$\text{sim}(w, w_t) = \cos(\theta) = \frac{v_w \cdot v_{w_t}}{|v_w| \times |v_{w_t}|}. \tag{18}$$

Similar to the SLDA model, the HME-LDA will set the parameter α_d separately for each document (each sentence in the SLDA model) based on the similarity between the vocabulary and seed words. The biased parameters can be calculated by equation (19).

$$\alpha_d = \frac{\sum_{i=1}^{N_d} \text{sim}(w_{d,i}, w_t)}{\sum_{t'}^{T} \sum_{i}^{N_d} \text{sim}\left(w_{d,i}, w_t'\right)} \times \alpha_{\text{base}}. \tag{19}$$

In equation (19), N_d is the number of all words in the current sentence, T is the number of topics, $w_{d,i}$ is the i-th word in the current sentence, and t is the seed word.

Based on the similarity between the vocabulary and seed words, the SLDA model will set parameters separately for each topic. These biased parameters can be calculated by equation (20), equation (21), and equation (22).

$$\beta_{t,A} = \text{sim}(w, w_t) \times \beta_{\text{base}}, \tag{20}$$

$$\beta_{t,P} = \text{sim}(w, w_{t,P}) \times \beta_{\text{base}}, \tag{21}$$

$$\beta_{t,N} = \text{sim}(w, w_{t,N}) \times \beta_{\text{base}}. \tag{22}$$

Different from SLDA, HME-LDA introduces the

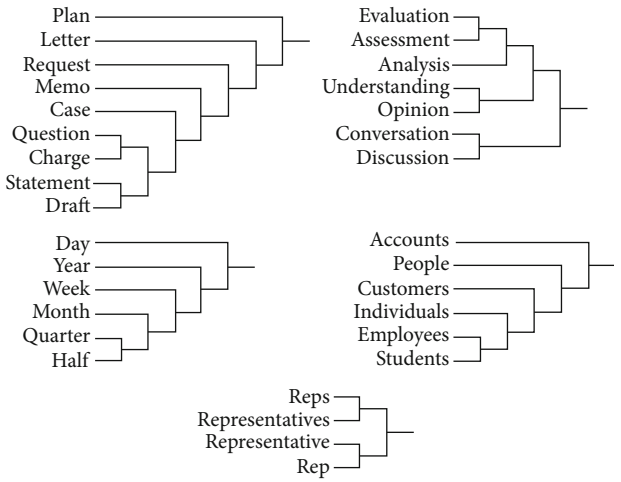

Figure 7: The results presentation of hierarchical clustering methods.

parameter δ to control the sentiment polarity of each sentence. The parameter $\delta_{d,q}$ can be calculated by equation (23).

$$\delta_{d,q} = \frac{\sum_i^{N_d} \mathrm{sim}\left(w_{d,i}, w_{t,q}\right)}{\sum_{q' \in \{P,N\}} \sum_i^{N_d} \mathrm{sim}\left(w_{d,i}, w_{t,q'}\right)} \times \delta_{\mathrm{base}}. \tag{23}$$

Similar to SLDA, HME-LDA uses the method of Gibbs sampling for solving. And the variables involved in the solution process have the same meanings as those in Tables 1 and 2.

Then, we use equation (24) to sample the topic of each sentence.

$$p\left(z_{d,n} = t \mid z_{\neg d,n}, y_{\neg d,n}, v_{\neg d,n}, \cdot\right) \propto \frac{n_{w_{d,n}}^{t,A} + \beta_{w_{d,n}}^{t,A}}{\sum_v^V \left(n_v^{t,A} + \beta_v^{t,A}\right)} \times \frac{n_{w_{d,n}}^{t,P} + \beta_{w_{d,n}}^{t,P}}{\sum_v^V \left(n_v^{t,P} + \beta_v^{t,P}\right)} \times \frac{n_{w_{d,n}}^{t,N} + \beta_{w_{d,n}}^{t,N}}{\sum_v^V \left(n_v^{t,N} + \beta_v^{t,N}\right)} \times \left(n_{d,t} + \alpha_{d,t}\right). \tag{24}$$

Equation (25) and equation (26) are used to sample $y_{d,n}$ and $v_{d,n}$,

$$p\left(y_{d,n} = u \mid z_{d,n} = t, \cdot\right) \propto \frac{n_{w_{d,n}}^{t,u} + \beta_{w_{d,n}}^{t,u}}{\sum_v^V \left(n_v^{t,u} + \beta_v^{t,u}\right)} \times \frac{e^{\lambda_u \times f_{d,n}}}{\sum_{u' \in \{A,O\}} e^{\lambda_{u'} \times f_{d,n}}}, u \in \{A, O\}, \tag{25}$$

$$p\left(v_{d,n} = q \mid z_{d,n} = t, \cdot\right) \propto \frac{n_{w_{d,n}}^{t,q} + \beta_{w_{d,n}}^{t,q}}{\sum_v^V \left(n_v^{t,q} + \beta_v^{t,q}\right)} \times \left(n_{d,q} + \delta_{d,q}\right). \tag{26}$$

In the corpus, the approximate probability of the topic t in sentence d can be calculated by equation (27),

$$\theta_d^t = \frac{n_{d,t} + \alpha_{d,t}}{n_d + \sum_{t'}^T \alpha_{d,t'}}. \tag{27}$$

With t as the topic, the approximate probability that the word $w_{d,n}$ is the opinion target can be calculated by equation (28),

$$\varphi_{w_{d,n}}^{t,A} = \frac{n_{w_{d,n}}^{t,A} + \beta_{w_{d,n}}^{t,A}}{\sum_v^V \left(n_v^{t,A} + \beta_v^{t,A}\right)}. \tag{28}$$

With t as the topic, the approximate probability that the word $w_{d,n}$ is the positive opinion word can be calculated by equation (29),

$$\varphi_{w_{d,n}}^{t,P} = \frac{n_{w_{d,n}}^{t,P} + \beta_{w_{d,n}}^{t,P}}{\sum_v^V \left(n_v^{t,P} + \beta_v^{t,P}\right)}. \tag{29}$$

With t as the topic, the approximate probability that the word $w_{d,n}$ is the negative opinion word can be calculated by equation (30),

$$\varphi_{w_{d,n}}^{t,N} = \frac{n_{w_{d,n}}^{t,N} + \beta_{w_{d,n}}^{t,N}}{\sum_v^V \left(n_v^{t,N} + \beta_v^{t,N}\right)}. \tag{30}$$

3. Results and Analysis

This chapter mainly trains the SLDA model in Section 2.1 and the HME-LDA model in Section 2.2 and, respectively, uses the above two models to extract the opinion targets and opinion review words on the *Restaurant* English data set, and then, this chapter will verify the feasibility of the models and analyse the experimental results.

3.1. Experimental Data Set. The data set of the experiment is from SemEval2016ABSA and Yelp. The original data of reviews on Yelp includes restaurants and hotels, so it needs to be screened. The test data is from the task B of SemEval2016ABSA. The test part of this experiment only pays attention to the classification of aspect-based opinion words, so the original test data need to be processed.

3.1.1. Original Data Set. SemEvel provides the training set and test set of the Restaurant reviews in XML format, where the size of the training set is 737 KB and the size of test set is 264 KB. The label structure of a sentence is shown in Figure 8.

The content in the label text is the original sentence, the attribute target in the label opinion is the opinion target, the category is the aspect category of reviews, and the polarity describes the sentiment polarity of the target. The model input proposed in this paper is the plain text without annotated information. The task of this paper is to extract the opinion target words and opinion sentiment opinion words and judge the polarity of the sentiment word. Here, the content in the text label needs to be extracted and added to the

```
<sentence id="1004293:1">
    <text>We, there were four of us, arrived at noon - the place was empty - and the
    staff acted like we were imposing on them and they were very rude.</text>
    <Opinions>
        <Opinion target="staff" category="SERVICE#GENERAL" polarity="negative" from="75"
        to="80"/>
    </Opinions>
</sentence>
```

Figure 8: The format of the SemEval data set.

corpus, and the content of target and category is extracted as test data. When verifying the experimental results based on the evaluation indicators, we only take the opinion target into account.

The size of the original data in Yelp is 231.2 MB, which contains lots of useless fields and covers a wide range of fields, including *Restaurant*, *Hotel*, and *Wine Bar*. There is a field, *business_categories*, which represents the realm of the review, in the Yelp data set. Therefore, the data can be filtered through the field above. The Yelp data is used to provide additional training sets and can be used to train Word2Vec and hierarchical clustering models to provide automatic annotated training sets for the maximum entropy model in HME-LDA.

3.1.2. Data Preprocessing. The Restaurant data set in SemEval is in XML format, while the format of Yelp data is CSV. Therefore, it is necessary to extract the data needed for this experiment from the files with two formats and carry out unified numbering. In the *Restaurant* data set, all sentences are annotated with text, so the sentences can be directly obtained from the labels. In the Yelp data set, the field, *business_categories*, stores the category of the review, and if the review contains the restaurant word, it is a restaurant-related review. The review of Yelp contains multiple sentences, so it can be split into multiple sentences by punctuations. Finally, the two txt files are used to store the extracted results, and one line in the file is a sentence. The xml.dom package and the csv package in python are used in the process. The amount of data finally extracted is shown in Table 3.

Repetitive words are removed from the statistics of the number of words, and no repetition words are removed in the statistics of the average length of sentences. When words are extracted from the sentences, they are separated by spaces and punctuation marks. Due to the need for additional tools for word type reduction, the word type reduction is not carried out here. Therefore, when counting the number of words, both the different tenses and the singular and plural forms of the same word are taken into account, resulting in the final number of words is larger than the fact.

There is additional annotated information in the SemEval corpus. The opinion target word and the aspect category of the word in the *Restaurant* review field can be extracted from the annotated information. The *opinion* label in the original xml file is extracted by python's xml.dom package, then the *attributes*, *target*, and *category* are extracted from the opinion label and make statistics. The statistical results are shown in Table 4.

Figure 9 generated by statistics in the training corpus of SimEval shows that the review category, *food*, accounts for a large proportion, while the review categories, *location* and *drinks*, account for a small proportion. The review category, *restaurant*, which contains a lot of semantics, is not usually the object of extraction. In this chapter, we only take the review categories, *food*, *service*, and *ambience* into account.

3.1.3. The Construction of the Experimental Data Set. The SemEval2016Restaurant review set is used in both the experimental data set and the test set. In the HME-LDA model, additional Yelp data sets are needed to train the hierarchical classification model and the Word2Vec model, and the amount of additional data provided will affect the final results as well. Therefore, the Yelp data set is divided into four training sets according to the number of sentences, which are shown in Table 5.

3.2. The Main Evaluation Indicators

3.2.1. The Evaluating Indicators of the Aspect Review Category of Sentences. Referring to the evaluation methods of Zhao et al. [22], this paper chooses accuracy P, recall R, and their harmonic value $F1$ as verification indicators, which are shown in equation (31), equation (32), and equation (33). In the MaxEnt-LDA [22] model, the topic of a sentence is undefined. According to the distribution of the topic words, the topics should be set to *food*, *service*, and *ambience* manually. Then, the sentence should be set to the corresponding topic according to the probability of the words appearing in each topic. The SLDA model proposed in this paper, as well as the HME-LDA model that has a fewer dependence on the language than SLDA, treats the seed word as the topic word and the aspect category of reviews. Therefore, the topic of a sentence can be determined directly by the distribution in the model.

$$P = \frac{\text{The number of correct predictions}}{\text{The total number of projections}}, \tag{31}$$

$$R = \frac{\text{The number of correct predictions}}{\text{The total correct number}}, \tag{32}$$

$$F1 = \frac{2 \times P \times R}{P + R}. \tag{33}$$

3.2.2. The Evaluation Indicators of the Opinion Target. In the models proposed in this paper, the distribution information of the opinion target based on the specific topic is stored in $\varphi_{t,A}$. The $\varphi_{t,A}$ preserves the word probability based on the

Table 3: The information of the experimental data set.

Index	SemEval	Yelp
The number of reviews	2000	85000
The number of words	3373	67066
The average length of sentence	13	13

Table 4: The information of the annotating data.

The aspect category of reviews	The number of sentences	The number of words before repetition	The number of words after repetition
Food	952	952	420
Restaurant	258	258	90
Service	324	324	57
Drinks	96	96	51
Ambience	228	228	94
Location	22	22	9

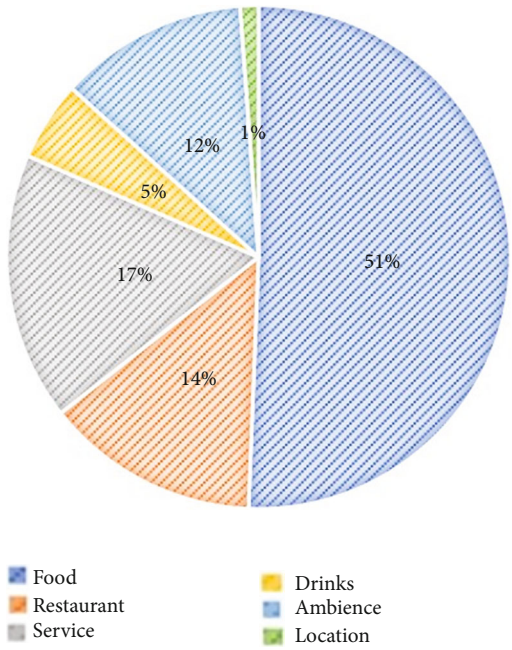

Figure 9: The proportion of review categories in the SemEval corpus.

Table 5: The information of the Yelp training set.

The name of the training sets	The number of sentences	The number of sentences
Yelp2K	2000	3809
Yelp4K	4000	5553
Yelp10K	10000	8948
Yelp20K	20000	12490

features of a topic model, and the same word may exist in different topics, so it is impossible to calculate the extracted accuracy and recall rate directly with $\varphi_{t,A}$. Referring to the scheme of Zhao et al. [22], n words with the highest probability of each topic in $\varphi_{t,A}$ are extracted and are treated as representatives, and then, their accuracy is calculated. In Section 1, the annotated data set contains information about the opinion targets and the categories of reviews they belong to. From this, the n words with the highest frequency of occurrence are selected as references for each review category. If n is 5, 10, 20, respectively, the accuracy rate $P_{t@n}$ can be calculated according to equation (31), where t is the topic number and n is the number of words taken. The average accuracy of extraction is expressed by $P_{t@n}$, which is calculated by equation (34).

$$P_{@n} = \frac{\sum_{t=0}^{T} P_{t@n}}{T}. \tag{34}$$

3.3. The Experimental Results and Analysis

3.3.1. The Setting of Experimental Parameters. In this experiment, the parameters related to the topic model are set as $\alpha_{\text{base}} = 50/T$, $\beta_{\text{base}} = 0.01$, and $\delta_{\text{base}} = T$, where T is the number of topics, and $T = 3$ in the subsequent experiment. In addition, the other one we need to set is the seed word and the cluster number of the hierarchical classification model. The seed word set is divided into two groups for comparison. One is the seed word set *A {food, service, ambience}*, and the other is seed word set *B {chicken, staff, atmosphere}*. The cluster number of the hierarchical classification model does not directly determine the number of categories. The cluster number is set to 200 according to experience.

The parameter setting of the comparison model MaxEnt-LDA [22] is the same as the original text. When training maximum entropy, the features of the model can be divided into three categories: word, part of speech, and part of speech plus word. The HME-LDA model proposed in this paper only uses the feature *word*. In the comparative experiment, the feature of the maximum entropy of the MaxEnt-LDA [22] model is chosen as the *word*.

3.3.2. The Influence of Seed Word Set on SLDA and HME-LDA. Figures 10 and 11 show the impact of seed words on the SLDA model and the HME-LDA model when they are, respectively, set to seed word set *A {food, service, ambience}* and seed word set *B {chicken, staff, atmosphere}*. When using the seed word set *A*, the model is recorded as SLDA-A, HME-LDA-A. When using the seed word set *B*, the model is recorded as SLDA-B, HME-LDA-B. Besides, the HME-LDA model additionally uses Yelp10K as the training set of the automatic annotated method.

It can be seen from the experimental results in Figure 10 that the selection of seed word sets has a greater impact on the SLDA model, while the impact on the HME-LDA model is not significant. Under the topics of *food* and *ambience*, the accuracy, recall rate and $F1$ value of the SLDA-A model are higher than those of SLDA-B, and when under the topic of *service*, the evaluation indexes above of SLDA-A model are

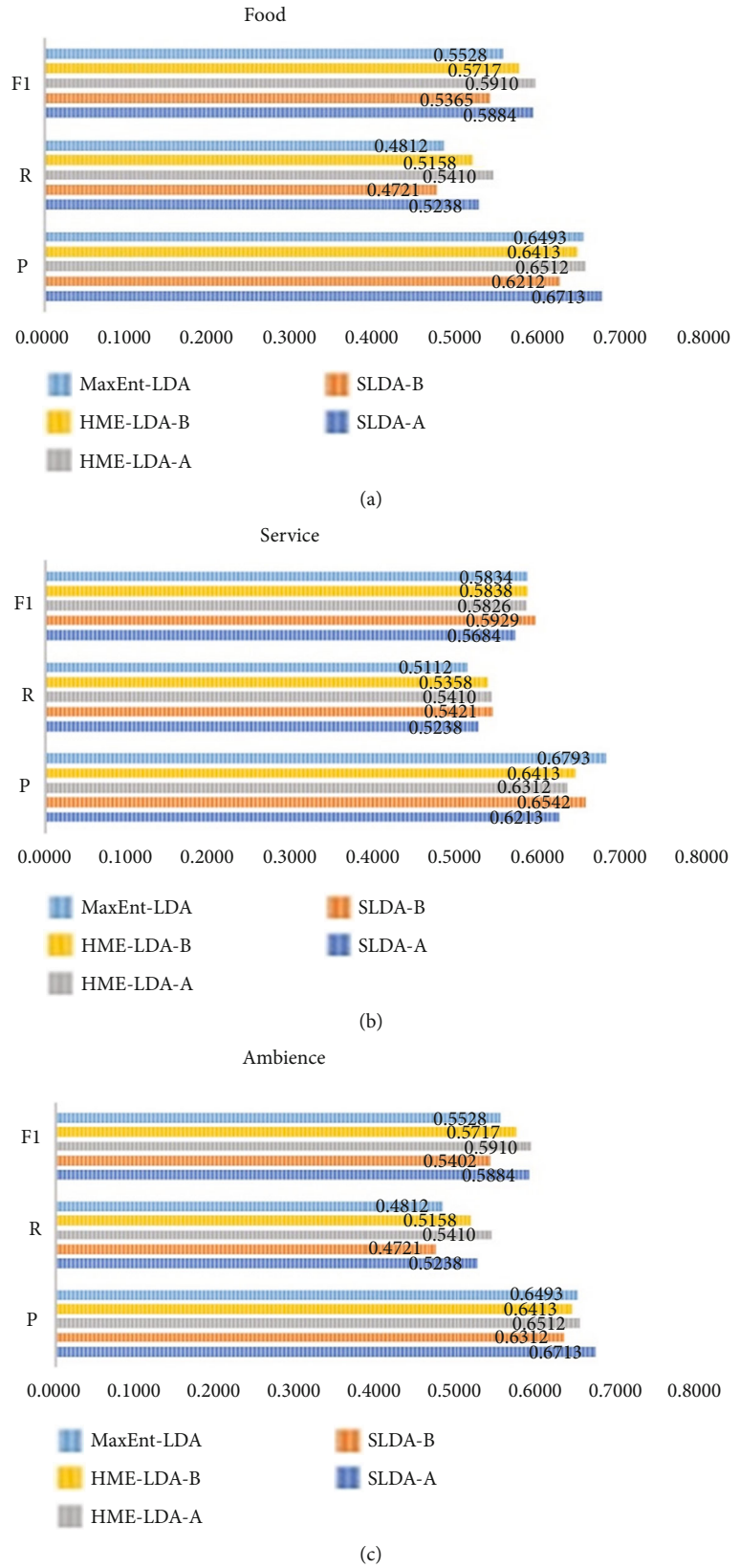

Figure 10: The influence of seed word sets on the extraction of sentence review categories: (a) evaluation index value of review category Food; (b) evaluation index value of review category Service; (c) evaluation index value of review category Ambience.

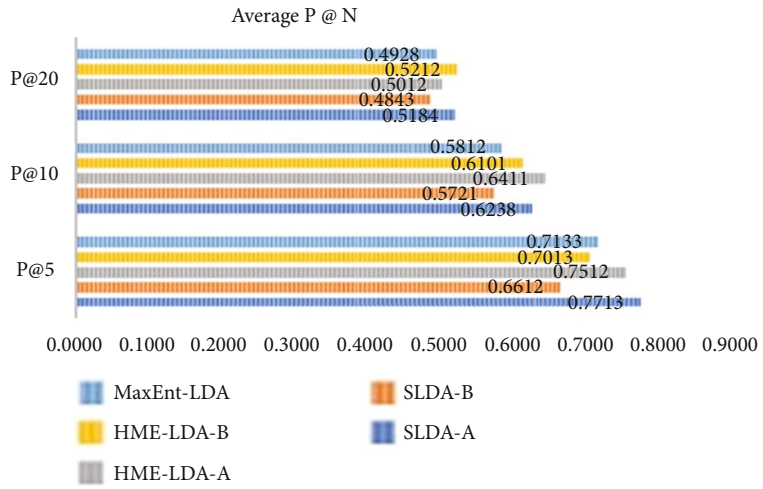

Figure 11: The influence of the seed word sets on P@N.

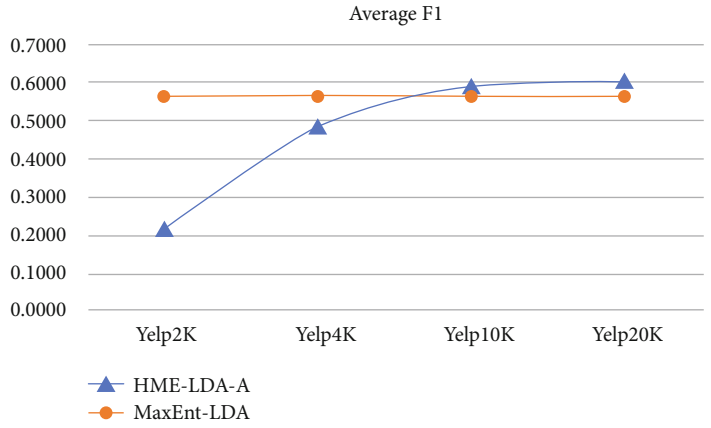

Figure 12: The F1 value of HME-LDA in training sets of different sizes.

lower than those of the SLDA-B model. This is related to the features of WordNet and SentiWordNet. In WordNet, the similarity between words is mainly calculated by the path between words. When the seed word *food* is used as the initial position, all branches only need to trace up to *food*. When a word in a branch of *food* is used as the seed word, the words of the other branches need to trace back to *food* first, and then searched down. This way makes the number of paths increase, which reduces the similarity of the words, thus affecting the results, while WordNet mainly preserves the concept of words and fails to perform knowledge reasoning, that is, it fails to infer the relationship between *waiter* and *service* to improve the similarity between them. When the seed word is replaced by *staff*, the explicit conceptual noun will make it easier to find noun words such as *waiter*, thus improving the effect of the SLDA. The eigenvector of the maximum entropy classifier in the HME-LDA model is taken as word order information, ignoring the conceptual problem of the word itself, so the choices of seed words have little effect on the final results. The evaluation indicators of SLDA-A and HME-LDA-A models are both slightly better than MaxEnt-LDA.

The results of Figure 11 show that the smaller the value of n, the higher the accuracy of the opinion target extraction in each model. And all models, namely, MaxEnt-LDA, SLDA, and HME-LDA, are suitable for extracting the opinion target words with the highest correlation, which is consistent with the comment habits of people. Most people only focus on a few aspect opinion targets of the comment objects, and the information people want to get from the comments is also based on their most concerned aspect. Similar to the results in Figure 11, when the seed word set is changed from A to B, the accuracy of the SLDA model decreases a lot, while the accuracy of the HME-LDA model decreases little, which has a great relationship with the word classification method of the SLDA. Meanwhile, it can be seen that SLDA-A and HME-LDA-A are slightly higher in the accuracy of the opinion target extraction than that of the MaxEnt-LDA.

3.3.3. The Influence of the Size of Training Set on HME-LDA. In the HME-LDA model, the effect of the maximum entropy model is related to the accuracy of automatically annotated data sets. The size of the training set that is used for training the hierarchical classification model will affect the accuracy

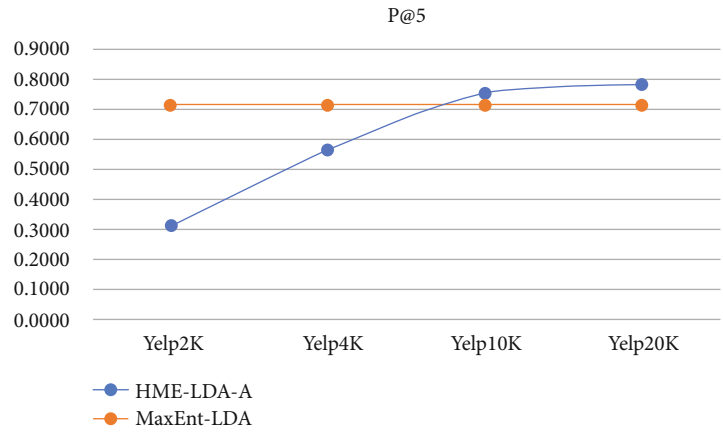

Figure 13: The P@5 value of HME-LDA in training sets of different sizes.

of automatic data annotation. Since it is difficult to calculate the accuracy of the automatic annotated data set, the impact of the training set size on the HME-LDA model is indicated by the evaluation indexes F1 and P@5 of the final HME-LDA model. Here, F1 takes the mean value of F1 under the three topics, and *P@5* indicates the mean value of the accuracy of the opinion target extraction in each review category when taking the top five values. In the experiment, the seed word is set to the seed word set *A {food*, *service*, *ambience}*. The experimental results are shown in Figures 12 and 13.

The value of the MaxEnt-LDA model is taken as a reference in the above figures. The Yelp data set is not used in the MaxEnt-LDA model, so the evaluation indicators of the MaxEnt-LDA model in these figures remain unchanged. With the increase of training set size, the effect of HEM-LDA is on the rise. When the training set reaches 10K, the effect of the HME-LDA model tends to be stable and slightly better than that of the MaxEnt-LDA model. The reason why the effect of the model is related to the size of the training set is because the larger the training set is, the more effective information it contains. Meanwhile, since the number of clusters in the hierarchical classification model is constant, when the amount of data is small, the number of words in each cluster is relatively small, and the accuracy of classification is relatively low. When the amount of data increases, the accuracy of classification will be improved relatively, which will affect the final effect of the model.

4. Conclusions

This paper mainly studies the methods to reduce the dependence of models on annotated data by focusing on the topic of aspect-based opinion mining. The unsupervised LDA topic model has good expansibility. Based on the LDA topic model, this paper introduces the two types of dictionary tools, WordNet and SentiWordNet, to propose the SLDA model. In order to further reduce the dependence on language tools, the maximum entropy model and the method of automatically annotating data are introduced to propose an optimized model HME-LDA. The experiments show that both the SLDA model and the HME-LDA model have good results on accuracy and recall rate without relying on the annotated data. Therefore, the two optimized models will give more detailed and accurate information for cryptocurrency investors in the blockchain to assist them in better decision-making.

Data Availability

Firstly, the Yelp data set used to support the findings of this study can be available from the https://www.yelp.com/dataset. Secondly, the SemEval2016ABSA data set used to support the findings of this study are included within the article: "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". Also, this dataset can be available from the http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Social Science Fund Planning Project of Ministry of Education of People's Republic of China "Research on Data Service and Guarantee for the Fourth Paradigm of Social Science" (20YJA870017).

References

[1] C. Song, H. Jung, and K. Chung, "Development of a medical big-data mining process using topic modeling," *Cluster Computing*, vol. 22, no. S1, pp. 1949–1958, 2019.

[2] L. Carson K-S, "Big data analysis and mining," in *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*, pp. 15–27, IGI Global, 2019.

[3] Q. Li, S. Li, S. Zhang, and J. Hu, "A Review of Text Corpus-Based Tourism Big Data Mining," *Applied Sciences*, vol. 9, no. 16, p. 3300, 2019.

[4] S. Lee, Y. Hyun, and M. J. Lee, "Groundwater potential mapping using data mining models of big data analysis in Goyang-si, South Korea," *Sustainability*, vol. 11, no. 6, article 1678, 2019.

[5] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, pp. 415–463, Springer, Boston, MA, 2013.

[6] K. Liu, L. Xu, and J. Zhao, *Opinion target extraction using word-based translation model*, pp. 1346–1356, 2012.

[7] Z. Chen, A. Mukherjee, and B. Liu, *Aspect extraction with automated prior knowledge learning*, 2014.

[8] H. Cheng, Z. Xie, Y. Shi, and N. Xiong, "Multi-step data prediction in wireless sensor networks based on one-dimensional CNN and bidirectional LSTM," *IEEE Access*, vol. 7, 2019.

[9] M. Pontiki, D. Galanis, H. Papageorgiou et al., *SemEval-2016 task 5: aspect based sentiment analysis*, pp. 19–30, 2016.

[10] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: aspect based sentiment analysis," in *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 27–35, 2014.

[11] Y. F. Zeng, T. Lan, and Z. F. Wu, "Bi-memory based attention model for aspect level sentiment classification," *Chinese Journal of Computers*, vol. 8, pp. 1845–1857, 2019.

[12] T. Chen, R. Xu, and X. Wang, "Improving Sentiment Analysis Via Sentence Type Classification Using BiLSTM-CRF and CNN," *Expert Systems with Applications*, vol. 72, pp. 221–230, 2017.

[13] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2018.

[14] T. Hofmann, "Probabilistic latent semantic indexing," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 211–218, 2017.

[15] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 993, 2013.

[16] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, *Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs*, 2007.

[17] F. Li, C. Han, M. Huang et al., *Structure-Aware Review Mining and Summarization*, vol. 2, pp. 653–661, 2010.

[18] D. Blei and J. McAuliffe, "Supervised topic models," *Advances in Neural Information Processing Systems*, vol. 3, 2010.

[19] D. Ramage, D. Hall, R. Nallapati, and C. Manning, *Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora*, vol. 1, pp. 248–256, 2009.

[20] H. Cheng, D. Feng, X. Shi, and C. Chen, "Data quality analysis and cleaning strategy for wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 61, 2018.

[21] H. Cheng, Z. Su, N. Xiong, and Y. Xiao, "Energy-efficient node scheduling algorithms for wireless sensor networks using Markov Random Field model," *Information Sciences*, vol. 329, pp. 461–477, 2016.

[22] W. Zhao, J. Jiang, H. Yan, and X. Li, *Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid*, pp. 56–65, 2010.