symmetry

MDPI

Article

WES-BTM: A Short Text-Based Topic Clustering Model

Jian Zhang [1,2], Weichao Gao [1,2] and Yanhe Jia [1,2,*]

[1] School of Economics and Management, Beijing Information Science and Technology University, Beijing 100192, China; zhangjian@bistu.edu.cn (J.Z.); 18888322343@163.com (W.G.)
[2] Beijing Key Lab of Green Development Decision Based on Big Data, Beijing 100192, China
* Correspondence: yhejia@bistu.edu.cn

Abstract: User comments often contain their most practical requirements. Using topic modeling of user comments, it is possible to classify and downscale text data, mine the information in user comments, and understand users' requirements and preferences. However, user comment texts are usually short and lack rich word frequency and contextual information with sparsity. The traditional topic model cannot model and analyze these short texts well. The biterm topic model (BTM), while solving the sparsity problem, suffers from accuracy and noise problems. In order to eliminate information barriers and further ensure information symmetry, a new topic clustering model, termed the word-embedding similarity-based BTM (WES-BTM), is proposed in this paper. The WES-BTM builds on the BTM by converting word pairs into word vectors and calculating their similarity to perform word pair filtering, which in turn improves clustering accuracy. Based on the experimental results using actual data, the WES-BTM outperforms the BTM, LDA, and NMF models in terms of topic coherence, perplexity, and Jensen–Shannon divergence. It is verified that the WES-BTM can effectively reduce noise and improve the quality of topic clustering. In this way, the information in user comments can be better mined.

Keywords: text mining; user comments; WES-BTM; visual analysis

Citation: Zhang, J.; Gao, W.; Jia, Y. WES-BTM: A Short Text-Based Topic Clustering Model. *Symmetry* **2023**, *15*, 1889. https://doi.org/10.3390/sym15101889

Academic Editor: Hsien-Chung Wu

Received: 19 August 2023
Revised: 27 September 2023
Accepted: 1 October 2023
Published: 9 October 2023

1. Introduction

Topic modeling is one of the most important methods in the field of text mining today and is important for understanding user needs and preferences [1]. The topic model analyses the structural and semantic information of text data by classifying and downgrading the text data to achieve a deep summary of the text [2]. The traditional LDA topic model is an unsupervised learning approach based on a probabilistic graphical model that generates polynomial distributions of topics a priori using Dirichlet [3]. However, when extracting topics from short texts such as user comments, the direct application of the LDA topic model will face the problem of severe sparsity. The BTM focuses on topic modeling for short texts and solves the problems of short text length and semantic sparsity by realizing word pair co-occurrence on the whole dataset [4]. Thus, the mining of hidden information in the short text corpus is realized. However, the BTM does not take into account the similarity between word pairs, which can lead to problems with accuracy and noise in the clustering results.

To solve the above problems, this paper introduces a WES-BTM (biterm topic model based on word embedding similarity). The WES-BTM uses word embeddings to convert word pairs into vector representations and computes the similarity between them for word pair filtering. The WES-BTM generates biterms from preprocessed data and generates word vectors with the help of pre-trained models. Cosine similarity is used to compare and calculate the similarity between two words to obtain the semantic relevance between them [5]. The parameter value γ was set and the screening range of similarity was obtained experimentally.

We compared the WES-BTM model with the BTM using user reviews from a typical manufacturing company, GREE, as a data source. We observed that the subject consistency scores of the WES-BTM are significantly higher compared with the BTM. It is verified that the WES-BTM can effectively reduce noise and improve the quality of topic clustering. Finally, we use the WES-BTM to mine the potential information in the text to construct thematic indicators for the analysis and identification of user needs, which enables enterprises to better understand user needs and further ensure the symmetry of information.

2. Literature Review

With the advent of the big data era, text data have exploded [6]. This has led to difficulty in using traditional data processing and analysis means to meet the current needs. In the face of this challenge, text mining has gradually become a mainstream method for text information analysis, as it is capable of extracting key information from text datasets [7]. Among the many techniques of text mining, topic modeling is particularly remarkable [8]. It is capable of detecting and outlining the intrinsic themes or topics of documents from large-scale documents. Topic modeling dissects the structural and semantic information in text data by classifying and downgrading the text data to achieve a deep summary overview of the text [2]. Various modeling techniques based on mathematical frameworks have been proposed in the research field of topic mining.

Deerwester et al. proposed the latent semantic analysis (LSA) model of LSA based on singular value decomposition (SVD), which discovers topic-based semantic relationships between text and words using matrix decomposition [9]. However, LSA lacks explicit probabilistic interpretation and requires large-scale matrix operations. For this reason, Hofmann proposed the probabilistic latent semantic analysis (PLSA) model, the core idea of PLSA is to build a probabilistic model for document–word item co-occurrence data [10]. Although PLSA solves some of the problems of LSA, the number of its parameters increases linearly with the number of documents, which may lead to overfitting. To solve this problem, Blei et al. [3] proposed a three-level probabilistic topic model LDA (latent Dirichlet allocation) based on semantic analysis research and then constructed a probabilistic generative model that generates polynomial distributions of topics with a Dirichlet prior. His proposed LDA topic model is widely used to mine hidden topics in text corpora.

LDA can handle large text data very well [11]. However, texts such as comments and tweets are usually short and contain only a few words [12]. This makes it difficult for LDA to estimate reliable topic distributions from short texts, and direct application will face the problem of serious sparsity. Many scholars have modeled topics by splicing short texts into long texts. Hong et al. [13] demonstrated that the length of a document can have a large impact on the effectiveness of a trained topic model and that model texts trained on summarized long texts have better performance. Zhao et al. [14] developed a Twitter-LDA model designed for short tweets, which finds topics from a representative sample of the entire Twitter sample to discover themes. Balikas et al. [15] found that grouping words in a sentence with the same theme can better handle the sparsity of short texts. In addition to these methods, some new technologies have emerged in the field of thematic modeling in recent years. Top2Vec is a novel method that utilizes vector embedding and clustering to discover topics from short texts [16]. BER Topic is another method that utilizes the powerful capabilities of the BERT language model for encoding and representing text [17]. By using pre-trained BERT models, BER Topic can capture rich semantic information from short texts and achieve more accurate topic modeling. Yan et al. [4] proposed a generative word pair theme model (BTM) to learn themes in short texts by directly modeling the generation of word pairs across the entire corpus. The BTM, by observing the co-occurrence patterns of word pairs across the entire dataset, can effectively provide a better model for short texts to infer topics. The BTM is also widely used in text clustering [18], topic mining [19], and hot topic detection [20], among others. In order to achieve better results, scholars have improved the BTM. Hu et al. selected words with specific parts of speech to form

binomials for the topic model, thereby improving the clustering effect of the model [21]. Huang et al. improved the dual-term topic model by introducing an embedded topic model to distinguish between noisy topics and potential topics [22]. Wu et al. proposed a short text clustering algorithm that combines the BTM and GloVe similarity linearly, thereby improving clustering performance [18].

In summary, the BTM is more suitable for short text topic mining, and more and more scholars process and analyze short text on the web using the BTM [23]. However, the traditional BTM does not take into account the correlation between word pairs while solving the sparsity problem of short text, which may add non-semantic heterogeneity and noise into the modeling process, thus affecting the effect of topic modeling. Therefore, in this paper, we propose a WES-BTM that uses the word2vec pre-training model for word vector processing and applies cosine similarity to calculate the similarity to realize word pair filtering, in order to reduce the noise and improve the quality of topic clustering, so as to be able to more accurately extract the topic content of short text.

3. Research Design

User reviews are typically short text data that contain valuable information about users' needs and attitudes [24]. In order to effectively extract topic information and potential information from text data, this paper proposes a WES-BTM for text mining, which can be summarized in four steps: (1) collecting product reviews from two e-commerce platforms, Tmall and JD.com. The data are then cleaned and word splitting is performed using Jieba, which is one of the most commonly used Chinese word splitting tools [25]; (2) creating biterms from the preprocessed data, vectorizing them using pre-trained models, and filtering out the biterms with low similarity by calculating their similarity (3) determining the optimal number of topics, modeling, and training to obtain topic clustering results; and (4) analyzing topic clustering results and visualizing them for further analysis. The specific process is shown in Figure 1.

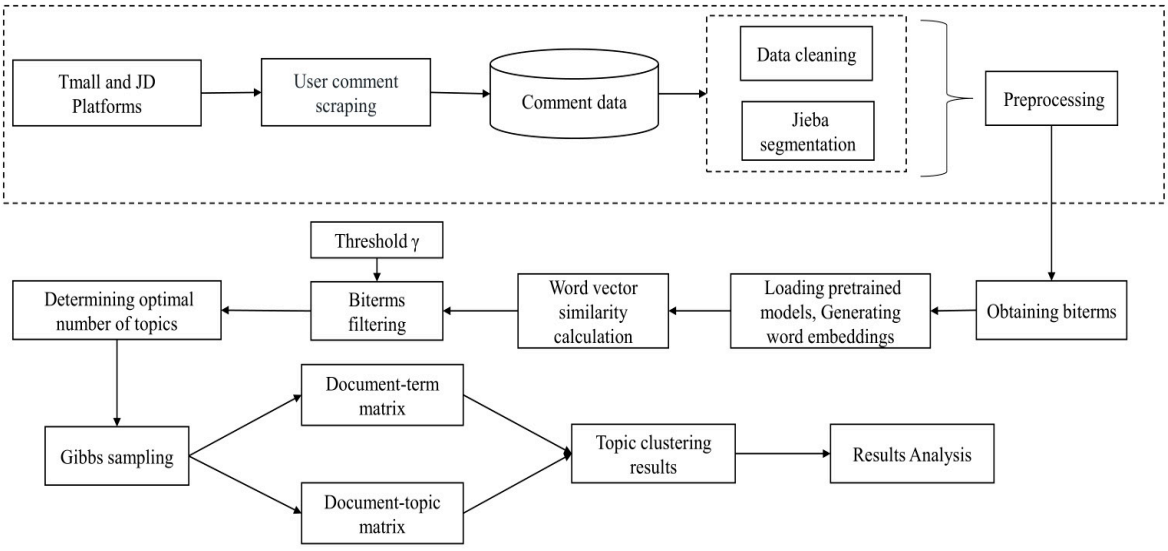

Figure 1. Process diagram.

3.1. Data Collection

With the development of e-commerce enterprises on the Internet, people's shopping habits have undergone tremendous changes, and online shopping has become the main way to purchase goods [26]. JD.com is China's largest self-operated e-commerce enterprise [27]. Tmall is the largest comprehensive e-commerce platform in terms of size in China [28]. The product reviews on these platforms can be used to obtain feedback from consumers.

GREE Electric Appliances is a prominent home appliance company in China, known for its large user base. GREE air conditioners are highly regarded for their excellent service

and high user satisfaction [29]. Analyzing the reviews of GREE enterprise can provide valuable insights into the actual feedback and opinions of users regarding a product. Hence, this research focuses on GREE Air Conditioning as the subject of study and collects user comments using a Python crawler, which will serve as the basis for further research.

3.2. WES-BTM

3.2.1. Biterms Screening

Product reviews are often in the form of short texts. Compared with ordinary documents, short texts lack word frequency and contextual information. Therefore, there is a serious sparsity problem when modeling with traditional topic models. The BTM represents topics as groups of related words, and their relevance is inferred from co-occurrence patterns [4]. By extracting all biterms from the entire corpus to construct training data, the sparsity problem of traditional topic models is solved. While modeling using the BTM, if the similarity between word pairs is low, it may lead to assigning word pairs belonging to the same topic to different topics. Then, the accuracy of the model is reduced. In addition, word pairs with low similarity may also increase the noise level of the model, thus making it more difficult for the model to distinguish topics. To compute the similarity between biterms, a pre-trained Word2Vec word embedding model is used [30]. Word2Vec is a method for capturing semantic information in natural language by learning word vectors from large-scale corpora [31]. This pre-trained model provides high-quality word embedding vectors. Then, the cosine similarity between biterms is calculated using the formula shown in Equation (1). Cosine similarity calculates word vector similarity, independent of vector length and direction, and is a simple and effective measure of data similarity. This method requires inputting two vectors representing biterms and returns a score between 0 and 1, where 1 indicates the highest similarity.

$$sim(v_i, v_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{\left|\vec{v}_i\right| \cdot \left|\vec{v}_j\right|} \tag{1}$$

3.2.2. Model Building

This article builds a WES-BTM based on the characteristics of user reviews for manufactured products. The WES-BTM probabilistic graphical model is shown specifically in Figure 2. The meanings of each parameter in the model are shown in Table 1.

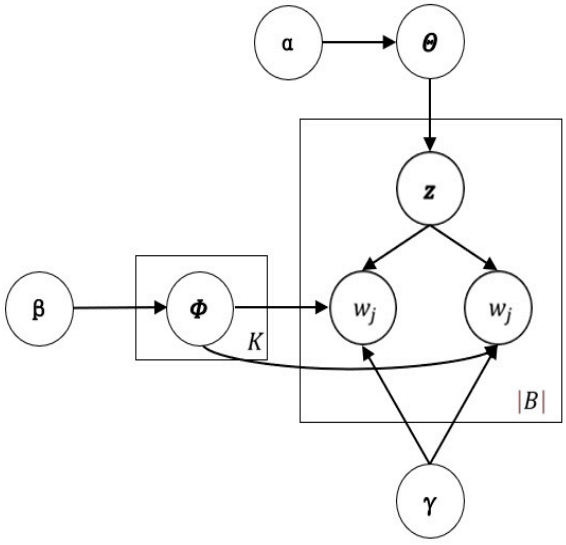

Figure 2. WES-BTM probabilistic graphical model.

Table 1. Meaning of each parameter in the WES-BTM.

Parameter	Meaning
α	Hyperparameter of Dirichlet distribution for θ
β	Hyperparameter of Dirichlet distribution for ϕ
ϕ	Topic-specific word distribution
θ	Topic distribution of the whole corpus
w_i, w_j	Biterms from the multinomial distribution ϕ of words
$\|B\|$	All biterms in the short text corpus
K	Dimensionality of topics
z	Topic selected from multinomial distribution θ of topics
r	Similarity threshold

The process of WES-BTM topic generation consists of three steps:

① For each specific topic z, calculate the topic distribution ϕ~Dir (β).

② Calculate the topic distribution θ across the entire corpus as $\theta \sim \mathrm{Dir}(\alpha)$.

③ Calculate the similarity between biterms and select them for screening.

For each selected word pair b in set B, perform the following operations: Suppose the selected biterm $b = (w_i, w_j)$: (1) Randomly draw a topic z from the entire collection of θ, that is, z~Muli (θ). (2) Randomly draw a pair of words w_i and w_j from topic z, that is, w_i, w_j~Muli (ϕ). Based on the above steps, the joint distribution probability of word pair b can be obtained, as shown in Equation (2), and the probability of the entire corpus can be obtained, as shown in Equation (3).

$$P(B) = \sum_z P(z)P(w_i|z)P(w_j|z) = \sum_z \theta_z \phi_{i|z} \phi_{j|z} \tag{2}$$

$$P(B) = \prod_{(i,j)} \sum_z \theta_z \phi_{i|z} \phi_{j|z} \tag{3}$$

Gibbs sampling simplifies the sampling process by decomposing the joint distribution into a series of conditional distributions, allowing us to sample one variable at a time. This sequential sampling method has high computational efficiency and is easy to converge. When modeling with the WES-BTM, using Gibbs sampling for global parameters θ and ϕ performs approximate inference as follows: Firstly, randomly select the initial state of a Markov chain [32]. Then, calculate the conditional distribution for each b = (w_i, w_j), where z_{-b} represents the topic allocation of all biterms except for b. By applying the chain rule to the joint probability of the entire data, the conditional probability can be obtained, as shown in Equation (4). Estimate the topic–word distribution ϕ and global topic distribution θ using word co-occurrences and counts of word–topic assignments as, shown in Equations (5) and (6).

$$P(z|z_{-b}, B, \alpha, \beta) \propto (n_z + \alpha) \frac{(n_{w|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2} \tag{4}$$

$$\theta_z = \frac{n_z + \alpha}{|B| + K_\alpha} \tag{5}$$

$$\phi_{\omega|z} = \frac{n_{\omega|z} + \beta}{(\sum_\omega n_{\omega|z} + M_\beta)} \tag{6}$$

In Equation (4), n_z denotes the number of occurrences of topic z in the corpus; $n_{w|z}$ denotes the number of occurrences of ω under the topic word z; and M is the number of distinct words in the corpus without considering repetition. In Equation (5), θ_z denotes the probability distribution of topic z in the corpus; K denotes the number of topics, and K_α denotes the hyperparameter α multiplied by the total number of topics. In Equation (6),

$\phi_{\omega|z}$ denotes the probability distribution of words ω under a given topic z and M_β denotes the hyperparameter β multiplied by the size of the vocabulary list.

3.3. Contrast Experiment

To verify the effectiveness of the model, we compared the WES-BTM with the following baseline models: the Nmf model, the LDA model, and the BTM. The Nmf model (non-negative matrix factorization) is a topic model method based on matrix decomposition. It assumes that the data in the text matrix are non-negative and decomposes the text matrix into two non-negative matrices, representing the relationship between document topic and topic word, respectively. The LDA model is a probability graph model used to learn hidden topic structures from text. It views documents as a mixture of topics and models the generation process of each word as sampling from the topic. The BTM is a topic model method for short text data. It learns topic distribution based on the co-occurrence relationship of all word pairs in the document.

This article uses topic consistency, perplexity, and Jensen–Shannon divergence to evaluate the effectiveness of the topic model.

Topic coherence is an indicator used to measure the consistency and interpretability of the topics generated using the topic model. It is based on the principle of word co-occurrence and evaluates the cohesion of a topic by calculating the similarity between different words. A higher degree of topic consistency means that the correlation between topic words is stronger and more explanatory. The formula is shown in Equation (7).

$$TC = \frac{2}{|T|(|T|-1)} \sum_{i=1}^{|T|} \sum_{j=i+1}^{|T|} sim(T_i, T_j) \tag{7}$$

where $|T|$ represents the number of topics and $sim(T_i, T_j)$ represents the similarity value obtained by calculating the correlation between the *i*-th and *j*-th topics. The specific formula for $sim(T_i, T_j)$ is shown in Equation (8):

$$sim(T_i, T_j) = \frac{\sum_{k,l \in Z} PMI(w_k, w_l)}{N} \tag{8}$$

where Z represents the word set composed of two topics, T_i and T_j, and N is the number of biterms. The specific formula for $PMI(w_k, w_l)$ is shown in Equation (9):

$$PMI(w_k, w_l) = \log \frac{P(w_k, w_l)}{P(w_k)P(w_l)} \tag{9}$$

where w_k and w_l stand for word or combination of words. k and l represent different indexes or numbers, which are used to identify different words. $P(w_k)$ and $P(w_l)$ represent the probability of words appearing in documents and $P(w_k, w_l)$ represents the probability of the two words appearing at the same time.

Perplexity is one of the commonly used evaluation indicators in topic models, which is used to measure the model's ability to predict new documents. The lower the perplexity, the stronger the model's ability to predict new documents. Perplexity can be understood as the uncertainty in the model's prediction of the next vocabulary for a given set of documents. Generally speaking, models with lower levels of perplexity are more able to accurately predict the next word.

$$Perplexity = \exp\left(-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right) \tag{10}$$

where W is a set of product review texts, M is the total number of review texts, $p(w_d)$ is the probability of a word in a review text, and N_d is the number of words in the review text.

Jensen–Shannon divergence is a measure of the difference between two probability distributions. In the topic model, we can use Jensen–Shannon divergence to measure the

similarity between the generated topic distribution and the real topic distribution. A lower Jensen–Shannon divergence means that the generated topic distribution is closer to the real distribution, indicating that the model performs better.

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{11}$$

where D_{KL} represents KL divergence, P and Q are two probability distributions, and M is the mean distribution of P and Q.

4. Result and Discussion

4.1. Data Collection and Preprocessing

We collected comments on a specific air conditioning product of GREE from the JD and Tmall platforms, obtaining 1000 comments from JD and 1980 comments from Tmall, totaling 2980 comments. We removed duplicate values from the crawled data, merged the initial and additional comments from users to ensure the authenticity and integrity of the comments., and eliminated phrases with a character length of less than four. The Harbin Institute of Technology Stop Words List is currently the mainstream stop word list [33]. In this paper, we use the Harbin Institute of Technology stop word list as the base stop word list, and add "GREE", "air conditioner", and other words that are not meaningful to the topic analysis. The stop words are excluded from the comments. Since a word is the smallest unit of writing in Chinese text, there is no obvious punctuation to distinguish between words. In the information processing of text, words are the components that can minimally represent the meaning of text. Jieba word segmentation tools are one of the most common Chinese word segmentation tools that provide various functions such as Chinese word segmentation and keyword extraction [34]. In this article, we use Jieba word segmentation tools to split the obtained product reviews for subsequent experiments.

4.2. Biterm Similarity Calculation

Word2Vec is a method for capturing semantic information in natural language by learning word vectors from large-scale corpora [31]. With the help of word vectors generated using the pre-trained model, we use the cosine similarity method to calculate the similarity between two words to obtain the semantic relevance between them. Some of the results obtained from the calculation are shown in Table 2. When "Door-to-Door" and "Delivered to home" appear together, it can be inferred that the topic category is more related to on-site maintenance. However, "Rest" and "Online shopping" cannot provide clear information about the topic. Therefore, it is necessary to remove biterms with low similarity using screening.

Table 2. Partial biterm similarity.

Word1	Word2	Similarity
Rest	Online shopping	0.099184
One-stop	Full process	0.46240
Door-to-door	Delivered to home	0.44340
Reliable	Cute	0.29023
Price Comparison	Cheap	0.59638
Impact	Repair	0.20116

4.3. Parameter Setting

To investigate the impact of parameter γ on model performance, this paper measures the clustering topic quality with TC value [35].

$$C(z; V^{(z)}) = \sum_{t=2}^{T}\sum_{l=1}^{t} \log \frac{D(v_m{}^{(z)}, v_l{}^{(z)}) + 1}{D(v_l{}^{(z)})} \tag{12}$$

where $D(v)$ is the document frequency of word type v, $D(v, v')$ is the common document frequency of word type v, and $V^{(t)} = \left(V_1^{(t)} \cdots V_m^{(t)}\right)$ is the possible word in the list topic T with the most M.

By calculating the semantic relevance of the topics, we evaluate the effectiveness of the topic model and improve the performance of the topic model. In order to select a suitable threshold value γ, the TC values of the BTM with γ = 0.1, 0.2, 0.3, and 0.4 and no improvement were respectively calculated under different numbers of topics.

Before training, various parameters were set. After conducting experiments and evaluating performance, we selected the following parameter values: α = 50/K, β = 0.01, and iterations = 200. Regarding the distribution of topics in documents, α determines the sparsity of topics. A smaller α value results in a sparser topic distribution, where the document focuses more on a few topics. Conversely, a larger α value leads to a more uniform topic distribution, with the document covering multiple topics more evenly. By setting α to 50/K, we can achieve a sparser topic distribution in large-scale problems, enabling more topics to concentrate on a few words and enhancing the model's ability to recognize keywords. β serves as a hyperparameter controlling the sparsity of the topic–word distribution. A smaller β value helps the model to better discover and distinguish keywords across different topics, thus improving the accuracy of the topic model. Conversely, a larger β value helps capture more general topics and common vocabulary. The experiment found that β = 0.01 yields better results. Regarding the selection of the iteration frequency, conducting 200 iterations allows us to fully leverage the corpus information in most scenarios. This process gradually stabilizes the model and yields a reliable topic distribution. Thus, considering the experimental results and performance evaluation, we chose the parameter values α = 50/K and β = 0.01 and iterations = 200 to obtain a model with a sparse topic distribution, high accuracy, and stability for large-scale problems. The results are shown in Table 3. The higher the TC value, the stronger the correlation between topics, which means a better topic model. When γ is set to 0.2, the results are better at multiple numbers of topics, and all experimental results are better than those without setting thresholds. So, in this paper, the γ value is set to 0.2.

Table 3. TC values under different numbers of topics.

Threshold γ	K = 5	K = 10	K = 15	K = 20	K = 25
0	−19.2738	−107.27	−268.705	−493.265	−810.569
0.1	−20.4315	−109.156	−261.272	−479.804	−797.129
0.2	−19.1881	−104.833	−265.059	−488.35	−775.682
0.3	−21.9134	−110.502	−275.766	−490.324	−779.5
0.4	−22.2633	−112.257	−266.203	−483.858	−744.105

4.4. Optimal Number of Topics

As the BTM belongs to unsupervised learning, the number of topics needs to be determined before modeling. Perplexity is an important indicator based on information entropy to determine the optimal number of topics [36]. Topic perplexity is an effective method to measure whether the number of topics is reasonable. Therefore, to select an appropriate number of topics, this study calculates perplexity according to Equation (10).

We calculated the perplexity score. The results are shown in Figure 3. According to the perplexity calculation results, as the number of topics increases, the perplexity continuously decreases. When the number of topics is 10, the rate of decrease in perplexity changes. So, 10 is chosen as the optimal number of topics.

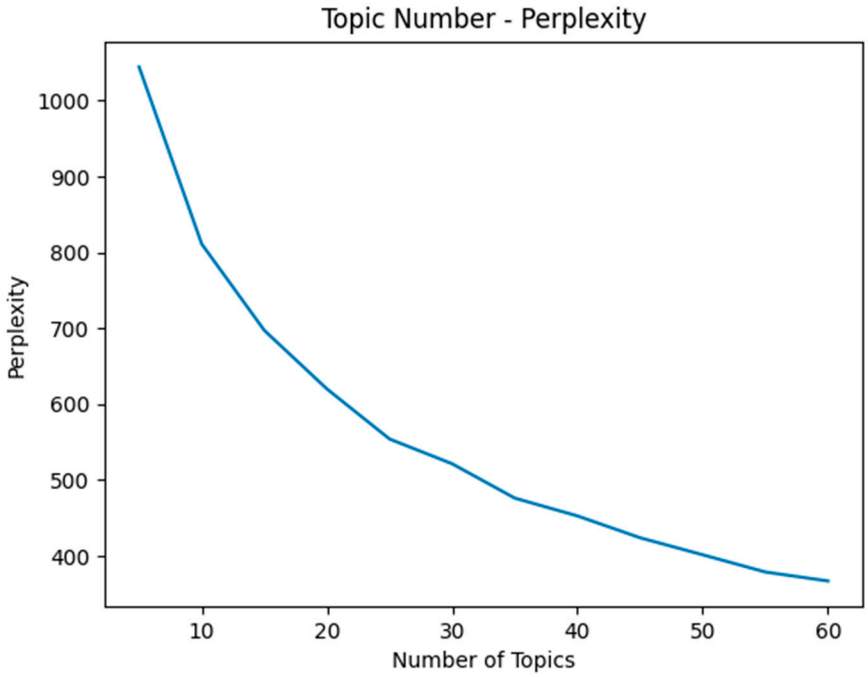

Figure 3. Perplexity score.

4.5. Topic Model Training

This study selected GREE enterprise product reviews as data acquisition channels on JD and Tmall platforms. After processing the comment data, a WES-BTM was constructed. Based on the calculation of perplexity, the number of topics is set to 10. The parameters are set to the following values: $\alpha = 50/K$ and $\beta = 0.01$. After training the topic model for product comment text, we obtained the generated keywords and topic distribution, as shown in Table 4.

Table 4. Topic keywords and distribution.

Topic	Keywords				
Topic 1	logistics	contact	delivery	arrival	shipment
Topic 2	return	charge	logistics	reply	refund
Topic 3	mode	function	brand	sleep	sleeping
Topic 4	quality	packaging	appearance	attitude	return
Topic 5	abnormal noise	solve	charge	repair	use
Topic 6	phone	reservation	contact	call	complaint
Topic 7	brand	quality	purchase	trust	like
Topic 8	personnel	brand	repair	old machine	solve
Topic 9	energy saving	quiet	energy efficiency	brand	heating efficiency
Topic 10	online	explanation	feedback	contact	handle

4.6. Results Analysis

4.6.1. Compared algorithms

Table 5 shows the comparison results of the WES-BTM and other baseline models. By computing coherence measures, we find that the WES-BTM achieves higher coherence scores, indicating its ability to generate more consistent and meaningful topics. Perplexity is an important metric for assessing model fit. We observe that the WES-BTM achieves significantly lower perplexity than the BTM, LDA, and NMF models, indicating its advantage in capturing patterns and structures within the data. We further evaluate the differences in topic distributions among the different methods by computing Jensen–Shannon. The results reveal that the Jensen–Shannon values between the topic distributions of the WES-BTM

and the true topic distributions are smaller, indicating that the WES-BTM better captures the underlying topic structure in the data.

Table 5. Results of comparative experiments.

Model \ Metric	Jensen–Shannon Divergence	Topic Coherence	Perplexity
NMF	0.732260575	69.29741217	1192.004004
LDA	0.726564254	69.99972076	1225.850567
BTM	0.679978854	72.62718216	816.7186215
WES-BTM	0.679495459	72.6664798	810.5492609

4.6.2. Topic Strength Analysis

The topic distribution of the user comments is shown in Table 6. In Table 6, the header row represents the topic, and the header column represents the comment number. The table displays the probability values of each comment's topic. The first comment discusses the excellent post-purchase services and the enthusiastic and meticulous repair personnel. Its topic is consistent with the semantics "on-line", "contact", and "processing" in Topic 10 and matches the probability calculation. In the second comment, the user describes the reliable air conditioning products and fast delivery, which is consistent with the semantics "brand", "quality", and "purchase" in Topic 7. In the third comment, the user praises the finely crafted and beautiful product with good quality, which is consistent with the semantics "brand", "quality", and "purchase" in Topic 7. The 2898th and 2899th comments describe the simple and good-looking appearance of the air conditioning products, good cooling effect, and energy-saving features, which are consistent with the semantics "energy-saving", "energy efficiency", and "brand" in Topic 9. Therefore, the probability of the comments belonging to Topic 9 is higher. We conducted document probability distribution statistics, and the number of comments belonging to each document topic is shown in Figure 4. Topic 9 has the most comments (1025), followed by Topic 7 (577), Topic 6 (277), Topic 3 (240), Topic 1 (220), Topic 8 (143), Topic 5 (129), Topic 2 (125), Topic 10 (100), and Topic 4 (63).

Table 6. Distribution of document topics.

Number \ Topic	1	2	3	4	5	6	7	8	9	10
1	0.02224	0	0.00205	0.00403	0.00487	0.0105	0.17302	0.00519	0.03042	0.74766
2	0.10231	0.00803	0.09877	0.01188	0.00761	0.06512	0.4701	0.00673	0.20856	0.02087
3	0.06096	0.00648	0.01792	0.00541	0.01029	0.01544	0.72332	0.00081	0.15876	0.0006
...										
2898	0.11855	0.01644	0.03072	0.00149	0.02322	0.05286	0.10575	0.0031	0.64223	0.00565
2899	0.04238	0.04128	0.03361	0.0012	0.00272	0.01534	0.02914	0.00877	0.75504	0.07053

4.6.3. Topic Visualization

Using the WES-BTM, we obtained the clustering results of 10 topics. Due to the high similarity between some topics, the clustering topics were further manually divided. The results were divided into four categories: logistics services, after-sales services, product quality and efficiency, and management and maintenance services, as shown in Figure 5.

The visualization of topic words generated using the WES-BTM is shown in Figure 6. Topic 1 is about logistics services. From the word cloud, we can see that the topic words for logistics services include "delivery", "logistics", "contact", and "arrival". Logistics is part of product services, and in order to accomplish logistics distribution, most companies cooperate with logistics companies. However, negative reviews from consumers about logistics services can also affect the products and the manufacturing companies themselves. Therefore, companies can improve consumer satisfaction with manufacturing industry products and services by integrating with the logistics industry.

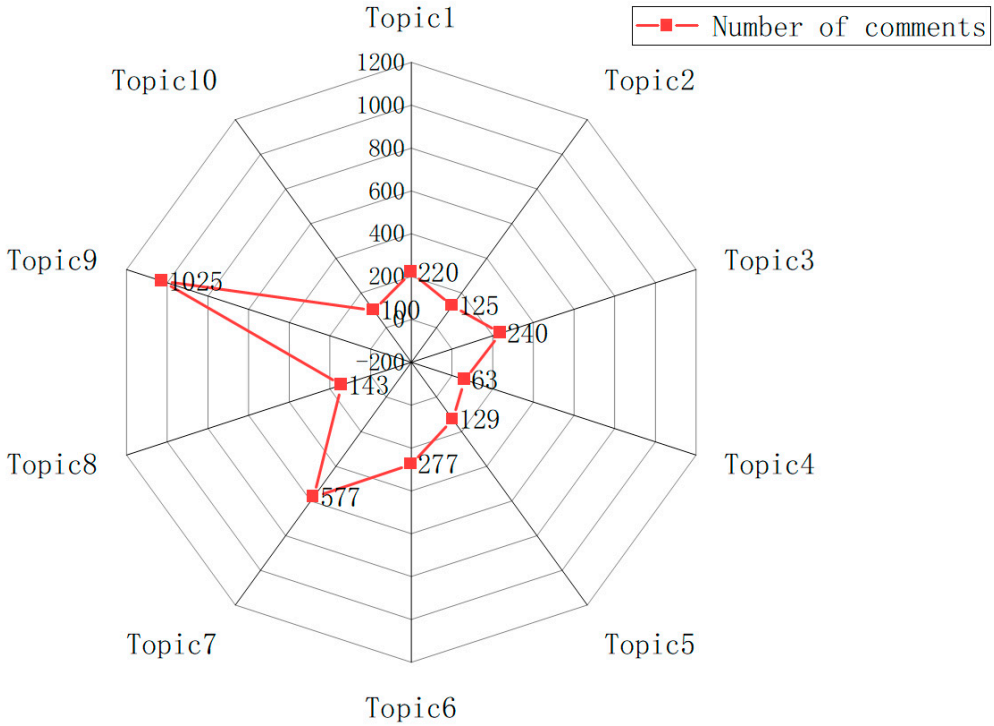

Figure 4. Distribution of the topic documents.

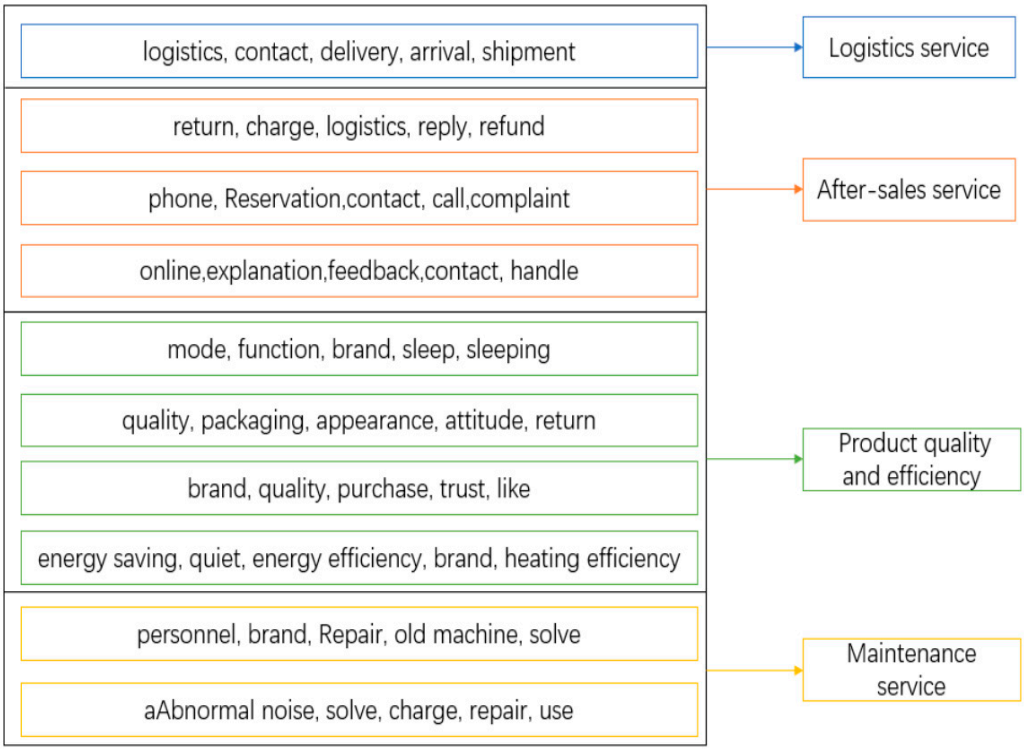

Figure 5. Results of topic segmentation.

Figure 6. Topic word cloud.

Topic 2 is about after-sales services, and its topic words include "contact", "phone", "handle", "return", and "charge". The primary goal of companies in promoting the service-oriented process is to increase customer satisfaction, thus enhancing the company's competitiveness and increasing its income. Contact refers to the interaction between the customer and the service team or support personnel. The customer can seek help by telephone, email, or online chat. Customer satisfaction is partly determined by the service of the service personnel. Improving the service awareness of service personnel and increasing service training is crucial to enhancing customer satisfaction with after-sales service.

Topic 3 is about product quality and efficiency. The topic keywords for product quality and efficiency include "brand", "quality", "mode", and "use". These keywords reflect customers' level of attention to the product itself. Customers' demands for products mainly include product quality and product efficacy. Manufacturing companies rely on the sale of high-quality and high-efficiency products. Customers attach great importance to the product itself, and for enterprises, focusing on product quality, improving product cost-effectiveness, and meeting customers' purchasing needs are the basis for enhancing their competitiveness. Products are the basis for customer consumption and the basis for the sustainable development of manufacturing enterprises. Because customers attach great importance to product quality at present, enterprises should pay attention to the products themselves while following the trend and reforming service-oriented development.

Topic 4 is about maintenance services. The topic keywords for maintenance services include "personnel", "solve", "machine", "contact", and "brand". After a manufacturing product is put into use, consumers often have a need for maintenance. The customer needs to communicate and contact the maintenance technician or maintenance department. Designing a consumer-preferred repair strategy is beneficial for manufacturers to consolidate their market position and enhance their market competitiveness. Product repair services determine to some extent the user's impression of the company and whether they will continue to choose to buy the brand's products.

5. Conclusions

A WES-BTM is proposed to address the shortcomings of short text with short length and sparse semantics, as well as existing models. This model obtains word vectors by

loading a Chinese Word2vec pre-training model. Next, the cosine similarity between word vectors is calculated and pairs of words with low similarity are filtered out. Finally, Gibbs sampling is performed on the selected word pairs to complete the training of the model. We use product reviews from JD and Tmall as data sources for the topic model. The results indicate that the WES-BTM performs well in multiple evaluation indicators such as topic coherence, perplexity, and Jensen–Shannon divergence. Finally, the results were further organized and visualized using comment data as an example.

Future research can further improve the WES-BTM to cope with more complex and diverse user comment data. In addition, this model can be applied to other fields, such as social media analysis, market research, etc., to explore more valuable user needs and attitude information, and provide a more accurate basis for enterprise decision-making.

Author Contributions: Conceptualization, J.Z. and W.G.; methodology, W.G. and Y.J.; software, W.G.; validation, J.Z.; formal analysis, Y.J.; investigation, W.G.; resources, J.Z.; data curation, Y.J.; writing—original draft preparation, W.G. and Y.J.; writing—review and editing, J.Z. and W.G.; visualization, Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China under Grant No. 2021YFF0901303.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lin, T.; Tian, W.; Mei, Q.; Cheng, H. The Dual-Sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text. In Proceedings of the 23rd International Conference on World Wide Web, New York, NY, USA, 7 April 2014; pp. 539–550.
2. Vulić, I.; De Smet, W.; Tang, J.; Moens, M.-F. Probabilistic Topic Modeling in Multilingual Settings: An Overview of Its Methodology and Applications. *Inf. Process. Manag.* **2015**, *51*, 111–147. [CrossRef]
3. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
4. Yan, X.; Guo, J.; Lan, Y.; Cheng, X. A Biterm Topic Model for Short Texts. In Proceedings of the 22nd International Conference on World Wide Web, ACM, Rio de Janeiro, Brazil, 13 May 2013; pp. 1445–1456.
5. Dehak, N.; Dehak, R.; Glass, J.; Reynolds, D.; Kenny, P. Cosine Similarity Scoring without Score Normalization Techniques. 2010. Available online: http://groups.csail.mit.edu/sls/publications/2010/Dehak_Odyssey.pdf (accessed on 1 December 2022).
6. Li, K.; Yan, D.; Liu, Y.; Zhu, Q. A Network-Based Feature Extraction Model for Imbalanced Text Data. *Expert Syst. Appl.* **2022**, *195*, 116600. [CrossRef]
7. Gnanavel, S.; Mani, V.; Sreekrishna, M.; Amshavalli, R.S.; Reta Gashu, Y.; Duraimurugan, N.; Srinivasa Rao, N. Rapid Text Retrieval and Analysis Supporting Latent Dirichlet Allocation Based on Probabilistic Models. *Mob. Inf. Syst.* **2022**, *2022*, e6028739. [CrossRef]
8. Qiu, L.; Yu, J. CLDA: An Effective Topic Model for Mining User Interest Preference under Big Data Background. *Complexity* **2018**, *2018*, 2503816. [CrossRef]
9. Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
10. Hofmann, T. Probabilistic Latent Semantic Analysis. *arXiv* **2013**, arXiv:1301.6705.
11. Anwar, W.; Bajwa, I.S.; Choudhary, M.A.; Ramzan, S. An Empirical Study on Forensic Analysis of Urdu Text Using LDA-Based Authorship Attribution. *IEEE Access* **2019**, *7*, 3224–3234. [CrossRef]
12. Tommasel, A.; Godoy, D. Short-Text Feature Construction and Selection in Social Media Data: A Survey. *Artif. Intell. Rev.* **2018**, *49*, 301–338. [CrossRef]
13. Hong, L.; Davison, B. Empirical Study of Topic Modeling in Twitter. In Proceedings of the SOMA 2010—Proceedings of the 1st Workshop on Social Media Analytics, New York, NY, USA, 25–28 July 2010. [CrossRef]
14. Zhao, W.X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; Li, X. Comparing Twitter and Traditional Media Using Topic Models. In Proceedings of the Advances in Information Retrieval, Dublin, Ireland, 18–21 April 2011; Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 338–349.
15. Balikas, G.; Amini, M.-R.; Clausel, M. On a Topic Model for Sentences. *ACM Sigir. Forum.* **2016**, 921–924.
16. Angelov, D. Top2Vec: Distributed Representations of Topics. *arXiv* **2020**, arXiv:2008.09470.
17. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794.
18. Wu, D.; Zhang, M.; Shen, C.; Huang, Z.; Gu, M. BTM and GloVe Similarity Linear Fusion-Based Short Text Clustering Algorithm for Microblog Hot Topic Discovery. *IEEE Access* **2020**, *8*, 32215–32225. [CrossRef]

19. Park, S.-H.; Song, A.-R.; Park, Y.-H.; Ihm, S.-Y. A Study on Bestseller Short Text Semantics Analysis Using Topic Model. *J. Image Cult. Contents* **2018**, *15*, 101–112. [CrossRef]
20. Niu, W.; Tan, W.; Jia, W. CS-BTM: A Semantics-Based Hot Topic Detection Method for Social Network. *Appl. Intell.* **2022**, *52*, 18187–18200. [CrossRef]
21. Hu, R.; Liu, J.; Wen, Y. SP-BTM: A Specific Part-of-Speech BTM for Service Clustering. In Proceedings of the 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Exeter, UK, 17–19 December 2020; pp. 1050–1057.
22. Huang, J.; Peng, M.; Li, P.; Hu, Z.; Xu, C. Improving Biterm Topic Model with Word Embeddings. *World Wide Web* **2020**, *23*, 3099–3124. [CrossRef]
23. Zhou, X.; Ouyang, J.; Li, X. Two Time-Efficient Gibbs Sampling Inference Algorithms for Biterm Topic Model. *Appl. Intell.* **2018**, *48*, 730–754. [CrossRef]
24. Zheng, X.; He, W.; Li, L. Distributed Representations Based Collaborative Filtering with Reviews. *Appl. Intell.* **2019**, *49*, 2623–2640. [CrossRef]
25. Fxsjy/Jieba: Jieba Chinese Word Segmentation. Available online: https://github.com/fxsjy/jieba (accessed on 18 September 2023).
26. Gao, J.; Zhang, W.; Guan, T.; Feng, Q. Evolutionary Game Study on Multi-Agent Collaboration of Digital Transformation in Service-Oriented Manufacturing Value Chain. *Electron. Commer. Res.* **2022**, 1–22. [CrossRef]
27. Li, R.; Jiang, Y.; Yang, W.; Tang, G.; Wang, S.; Ma, C.; He, W.; Xiong, X.; Xiao, Y.; Zhao, E.Y. From Semantic Retrieval to Pairwise Ranking: Applying Deep Learning in E-Commerce Search. In Proceedings of the Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Paris, France, 18 July 2019; pp. 1383–1384.
28. Xin, S.; Li, Z.; Zou, P.; Long, C.; Zhang, J.; Bu, J.; Zhou, J. ATNN: Adversarial Two-Tower Neural Network for New Item's Popularity Prediction in E-Commerce. In Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 19–22 April 2021; pp. 2499–2510.
29. Chen, Y.; Visnjic, I.; Parida, V.; Zhang, Z. On the Road to Digital Servitization—The (Dis)Continuous Interplay between Business Model and Digital Technology. *Int. J. Oper. Prod. Manag.* **2021**, *41*, 694–722. [CrossRef]
30. Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; Du, X. Analogical Reasoning on Chinese Morphological and Semantic Relations. *arXiv* **2018**, arXiv:1805.06504.
31. Zhang, D.; Xu, H.; Su, Z.; Xu, Y. Chinese Comments Sentiment Classification Based on Word2vec and SVMperf. *Expert Syst. Appl.* **2015**, *42*, 1857–1863. [CrossRef]
32. Bayesian Networks: Regenerative Gibbs Samplings: Communications in Statistics—Simulation and Computation: Vol 51, No 12. Available online: https://www.tandfonline.com/doi/abs/10.1080/03610918.2020.1839770?journalCode=lssp20 (accessed on 30 March 2023).
33. Cao, H.; Kang, J. Study on Improvement of Recommendation Algorithm Based on Emotional Polarity Classification. In Proceedings of the 2020 5th International Conference on Computer and Communication Systems (ICCCS), Shanghai, China, 15–18 May 2020; pp. 182–186.
34. Wang, X.; Wang, H.; Zhao, G.; Liu, Z.; Wu, H. ALBERT over Match-LSTM Network for Intelligent Questions Classification in Chinese. *Agronomy* **2021**, *11*, 1530. [CrossRef]
35. Mimno, D.; Wallach, H.M.; Talley, E.; Leenders, M.; McCallum, A. Optimizing Semantic Coherence in Topic Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Scotland, UK, 27–31 July 2011; pp. 262–272.
36. Ma, Y.; Xiang, Z.; Du, Q.; Fan, W. Effects of User-Provided Photos on Hotel Review Helpfulness: An Analytical Approach with Deep Leaning. *Int. J. Hosp. Manag.* **2018**, *71*, 120–131. [CrossRef]