


## Article

# Tourism Review Sentiment Classification Using a Bidirectional Recurrent Neural Network with an Attention Mechanism and Topic-Enriched Word Vectors

Qin Li <sup>1,2</sup>, Shaobo Li <sup>3,4,\*</sup> , Jie Hu <sup>3</sup>, Sen Zhang <sup>1,2</sup> and Jianjun Hu <sup>3,5,\*</sup>

<sup>1</sup> Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China; qinlee85@126.com (Q.L.); 20120061@git.edu.cn (S.Z.)

<sup>2</sup> University of Chinese Academy of Sciences, Beijing 100049, China

<sup>3</sup> School of Mechanical Engineering, Guizhou University, Guiyang 550025, China; jason.houu@gmail.com

<sup>4</sup> Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China

<sup>5</sup> Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

\* Correspondence: lishaobo@gzu.edu.cn (S.L.); jianjunh@cse.sc.edu (J.H.); Tel.: +1-803-777-7304 (J.H.)

Received: 14 August 2018; Accepted: 3 September 2018; Published: 17 September 2018



**Abstract:** Sentiment analysis of online tourist reviews is playing an increasingly important role in tourism. Accurately capturing the attitudes of tourists regarding different aspects of the scenic sites or the overall polarity of their online reviews is key to tourism analysis and application. However, the performances of current document sentiment analysis methods are not satisfactory as they either neglect the topics of the document or do not consider that not all words contribute equally to the meaning of the text. In this work, we propose a bidirectional gated recurrent unit neural network model (BiGRULA) for sentiment analysis by combining a topic model (lda2vec) and an attention mechanism. Lda2vec is used to discover all the main topics of review corpus, which are then used to enrich the word vector representation of words with context. The attention mechanism is used to learn to attribute different weights of the words to the overall meaning of the text. Experiments over 20 NewsGroup and IMDB datasets demonstrate the effectiveness of our model. Furthermore, we applied our model to hotel review data analysis, which allows us to get more coherent topics from these reviews and achieve good performance in sentiment classification.

**Keywords:** topic model; attention mechanism; lda2vec; BiGRU; sentiment classification; tourism review

## 1. Introduction

The availability of extensive tourism online reviews provides an unprecedented opportunity to analyze the emotions, preferences, feelings, and opinions expressed by visitors. Sentiment analysis is one of the major techniques for this purpose, which provides us insight into tourism services. Currently, a significant amount of research has been carried out on tourism analysis and applications based on sentiment analysis. Zheng [1] proposed a tourism destination recommender system by analyzing and quantifying users' sentiment tendency. Ren [2] proposed a topic-based sentiment analysis approach to measure online destination image. Li [3] designed a visual analytic system to analyze tourists' regional tendency and sentiment changes from user-generated content (UGC) data. Serna [4] analyzed the public bike share system in Spain to explore sustainable tourism through sentiment analysis of UGC. He [5] used sentiment analysis techniques to analyze online hotel reviews and to understand users' preferred hotel attributes or demands.

With an increasing demand for sentiment analysis of text data, a variety of research on improving the accuracy of document sentiment classification was carried out. The goal of document sentiment

classification is to assign emotional labels, such as positive, negative, neutral, and so on, to the document. A key problem in document sentiment classification is feature representation or selection. Many works were done using methods from simple  $n$ -grams to topic models, to the recently developed deep learning. For example, Tripathy [6] used CountVectorizer and term frequency-inverse document frequency (TF-IDF) to represent movie reviews and used various  $n$ -gram machine learning methods to classify them. Hu [7] explored the topical term description models to conduct document sentiment classification. Kalchbrenner [8] proposed a convolutional neural network for semantic modeling of sentences with dynamic  $k$ -max pooling. Lai [9] replaced traditional window-based neural networks with recurrent structures for text classification. However, these works either neglect the order of the sentences or neglect the global meaning of the document vectors.

In this paper, we present a new approach (BiGRULA) based on a bidirectional gated recurrent unit (BiGRU) neural network for sentiment classification, which combines the topic model, lda2vec, and an attention mechanism. Our model can be used as a feature extractor for texts. It uses the GRU as the basic model, which keeps the sequence order, and adapts gated mechanism to deal with the problem of long-term dependencies and gradient vanishing.

The main contributions of this paper are summarized as follows:

- (1) We proposed the BiGRULA recurrent neural network model with topic-enhanced word embedding and an attention mechanism for sentiment classification.
- (2) We evaluated and showed the advantage of using topic-enhanced word embedding based on the lda2vec model for document classification compared with other text representations.
- (3) We evaluated and compared the performance of BiGRULA for sentiment classification with other neural network models. Our algorithm achieved an accuracy of 89.4% with 3.0% improvement over the best of three baseline algorithms.
- (4) We applied our BiGRULA model to a real-world hotel review comment analysis and demonstrated its capability to extract meaningful topics from the reviews and to make accurate sentiment classification.

## 2. Related Work

Major progress was made recently in sentiment analysis, ranging from word embedding methods to recurrent neural networks. For example, word2vec [10] is one of the widely used word-embedding models and is used in a variety of applications related to text processing. However, it still has some limits. For example, it cannot solve the problem of polysemy, and the learned word vector cannot represent the global meaning. To address this limitation, latent Dirichlet allocation (LDA) [11] was proposed as a probabilistic topic model that can extract latent topics from documents. It describes the topic distribution of the documents and word distribution of the topic by probability distribution, and can represent a global rather than contextual relationship.

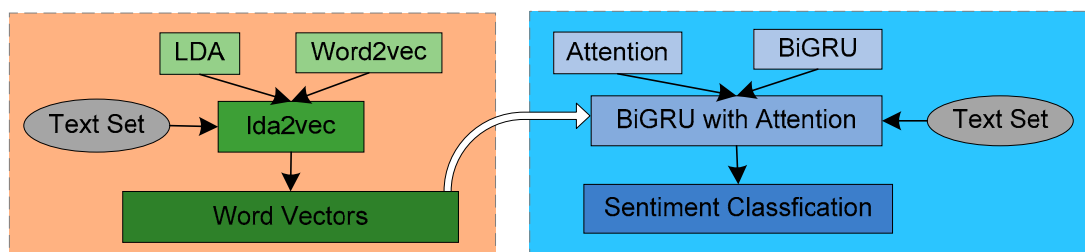
Based on these methods, a hybrid document feature extraction method was put forward [12]. This method uses the latent Dirichlet allocation and word2vec independently to train topic vectors, while the document vector is still the simple average of the word vectors. Liu et al. [13] proposed a topical word embedding (TWE) based on all the words and their topics. Compared to word2vec, it uses the topic of the words to predict the context and allows the same word to have different word vector expressions under different topics. Yao et al. [14] combined word2vec and LDA to mine coherent topics in documents. Zhang [15] learned from LDA to supervise the training of deep neural networks. All these works aim to combine word2vec with LDA, expecting to take advantage of both; however, they still cannot train a type of word vector that can both represent the local meaning of documents and explain the global meaning of topic distribution. In 2016, Moody [16] proposed a model named lda2vec by mixing Dirichlet topic models and word embedding. This model attempts to construct a context vector by adding the composition of a document vector and the word vector, which are all learned during the training process. It greatly improves the representation power of standard word vectors.

Another major research theme in sentiment analysis is how to compute the sentence and document vectors. Several works tried importing document topics and an attention mechanism into the neural network framework for document classification. Dieng [17] proposed TopicRNN, which integrates the merits of recurrent neural networks (RNNs) and the latent topic model to achieve long-range semantic dependency. Li [18] proposed a recurrent attentional topic model for document embedding based on a novel recurrent attentional Bayesian process. A feed-forward network with attention was suggested by Raffel [19], which is a well-known work using the attention mechanism.

Attention mechanism is a powerful technique for solving the problem of long-term memory for sequences. It is widely used in various tasks, especially with recurrent neural network models. Whatever the topic model or attention mechanism, they both aim to extract more informative features, which could be combined for achieving better sentiment analysis performance as we proved below. A combination of the attention mechanism with topic information was used in Reference [20] for text summarization based on a convolutional sequence-to-sequence model, which is different from the sentiment analysis task in this paper.

### 3. BiGRULA

We propose the BiGRULA document sentiment analysis model based on lda2vec and the attention mechanism. The overall architecture is shown in Figure 1. In the left part of this architecture, the lda2vec model, which is based on LDA and word2vec, is used to extract document topic-based word vector representation. It adds the context information to the word embedding. Through lda2vec, we can get the word vectors and the topics from text dataset. Then the topic-enhanced word vectors are used to encode the text set, which are then fed to the BiGRU recurrent neural network model with attention to get the document vectors and the classification model. Finally, the text documents are classified by this model.



**Figure 1.** The bidirectional gated recurrent unit neural network model (BiGRULA) framework for sentiment analysis.

#### 3.1. Lda2vec Architecture

We exploited the lda2vec algorithm as the topic feature extractor of the BiGRULA model. Lda2vec has the ability to extract topics from texts and to generate topic-adjusted word vectors, which makes these word vectors more interpretable by linking them to the topics. In other words, sparse word vectors in our model are enhanced with meaning by importing the interpretable document representation. The lda2vec model mainly takes advantage of the document representation and learns the topic weights of the documents by minimizing the objective function of the skip-gram negative sampling (SGNS). The procedure of the lda2vec model is illustrated in Figure 2. Details of the model are explained below.

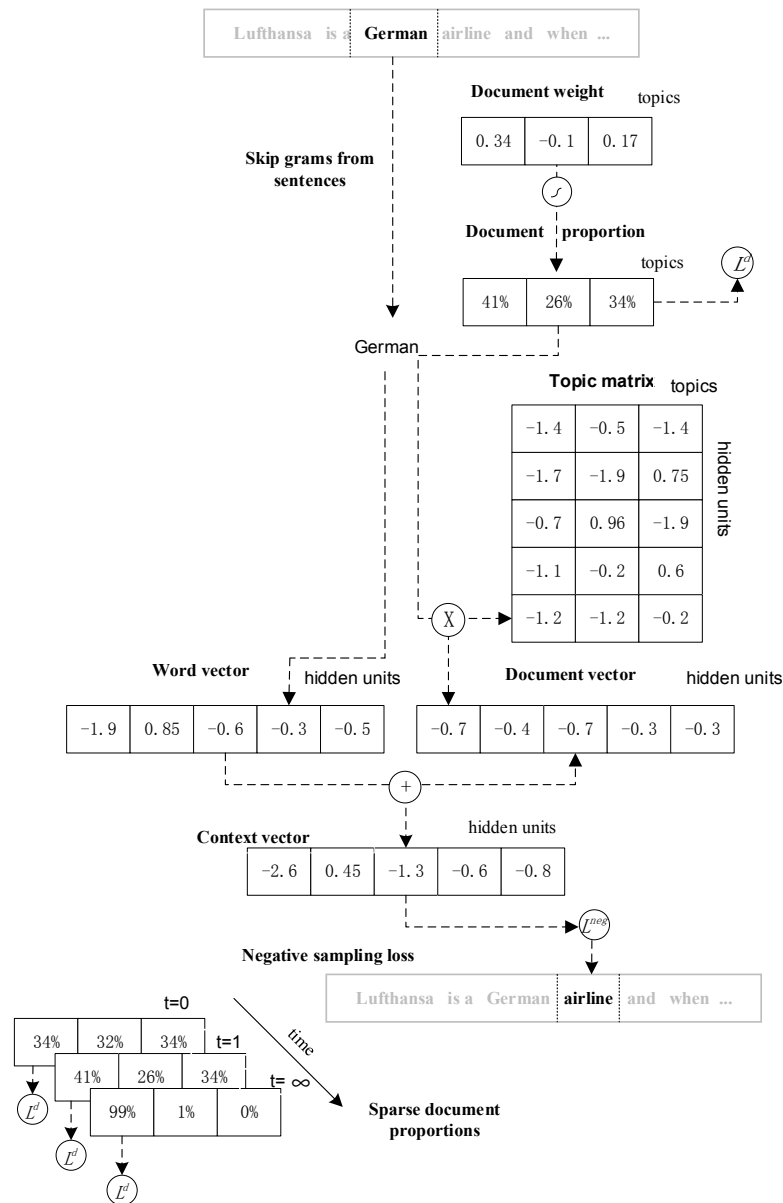


Figure 2. Workflow of lda2vec model for topic-enhanced word vector encoding [16].

### 3.1.1. Document Vector

Document vectors are used to represent the topic tendency of a document. We used lda2vec to obtain an interpretable representation of a traditional LDA to generate document vectors. In order to achieve this goal, a document vector was calculated as a weighted sum of topic vectors.

$$\vec{d}_j = \sum_{k=1}^n p_{jk} \vec{t}_k, \quad (1)$$

where  $p_{jk}$  denotes the weight of topic  $k$  on document  $j$ ,  $p_{jk} > 0$ ,  $\vec{t}_k$  denotes  $k$ -th topic vector, and  $n$  is number of topics. During the training process, the document vector was updated by these weights which were normalized to ensure  $\sum_k p_{jk} = 1$ . The document vector, word vector, and topic vector were all in the same vector space. To determine the specific meaning of a topic vector, we only need to compute the most similar words with the topic vector.

Finally, topic weights of the document were optimized using Dirichlet likelihood  $L^d$ .

$$L^d = \gamma \sum_{jk} (\alpha - 1) \log p_{jk}, \quad (2)$$

where  $\gamma$  denotes the strength of the  $L^d$  in the training process of lda2vec, and  $\alpha$  denotes a low concentration parameter. If  $\alpha$  is less than 1, topics will become sparse; if  $\alpha$  is equal to 1, Dirichlet will degenerate to uniform distribution and leads to poor consistency of topics; if  $\alpha$  is more than 1, the difference between topics will become small.

In the beginning, the weights of all topics are initialized to be the same. As the iterative training goes on, they become sparser and concentrate on one or a few topics.

### 3.1.2. Context Vectors

Context vectors are topic/context-enhanced word vectors used in our BiGRULA model. They were calculated as follows: firstly, given a pivot word in the text corpus, five target words in a moving window behind and after the pivot word were selected. This process was repeated across all the corpus. Then, the pivot word was used to predict the nearby target words. For example, if the pivot word is “red”, then the nearby words are probably predicted as “green” or “yellow”. Assuming that we know the document is about weather, the words nearby the pivot word “red” should be more likely to be predicted as “typhoon” or “heavy rain” or “high temperature”.

Context vectors are inspired by the meaningful word vector combination through addition and subtraction of word vectors such as “king – man + woman = queen”. Equally, if we add a word vector and the document vector together, the sum vector will, thus, capture long- and short-term themes. In our model, the context  $\vec{c}_i$  was defined as the addition of the pivot vector  $\vec{\omega}_i$  and the document vector  $\vec{d}_j$ .

$$\vec{c}_j = \vec{\omega}_j + \vec{d}_j, \quad (3)$$

where  $\vec{\omega}_j$  is the word vector of  $j$ -th word in the document.

### 3.1.3. SGNS (Skip-Gram Negative Sampling)

In our model, we used SGNS to jointly train the context vectors and topic-enhanced word vectors as shown in Figure 1. SGNS attempts to differentiate the target words from the negative samples which are randomly picked from a negative sampling pool. In SGNS, high-frequency words are selected as the negative samples with higher probability, while low-frequency words are less likely to be selected. Let  $\mu$  denote the word frequency normalized by the sample scale; then, the probability of the appearance of infrequent words is regulated and controlled by  $\mu^\beta$  where  $\beta$  is a smoothing parameter. When the target word is separated from the negative samples, the loss function  $L_{ij}^{neg}$  is minimized.

$$L_{ij}^{neg} = \log \sigma(\vec{c}_j \cdot \vec{\omega}_i) + \sum_{l=0}^n \log \sigma(-\vec{c}_j \cdot \vec{\omega}_l), \quad (4)$$

where  $\vec{\omega}_i$  denotes the word vector of the target word  $i$ , and  $\vec{\omega}_l$  denotes the word vector of negative sample  $l$ . To prevent self-adaption, we used dropout on the document vector and the pivot vector before they were normalized.

In the process of negative sampling, we need to remove high-frequency stop words to reduce the noise of the model. We used Equation (5) to calculate the probability of the word being canceled.

$$p(\omega_i) = 1 - \left( \sqrt{\frac{t}{f(\omega_i)}} + \frac{t}{f(\omega_i)} \right), \quad (5)$$

where  $t$  denotes the threshold, and  $f(\omega_i)$  denotes the frequency of word  $\omega_i$ .

### 3.1.4. Loss Function

The whole loss function was as follows:

$$L = L^d + \sum_{ij} L_{ij}^{neg}. \quad (6)$$

### 3.2. BiGRU with Attention Mechanism

In our BiGRULA model for sentiment analysis, we used the BiGRU model with an attention mechanism to build document vectors from a sequence of word vectors, which were then used to make classification for documents. Attention mechanism has two benefits: firstly, it helps the model get better performance; secondly, it provides a mechanism to assign different importance to different words in document classification. Next, we introduce the model of BiGRU and attention mechanism in detail. The architecture of the model is illustrated in Figure 3.

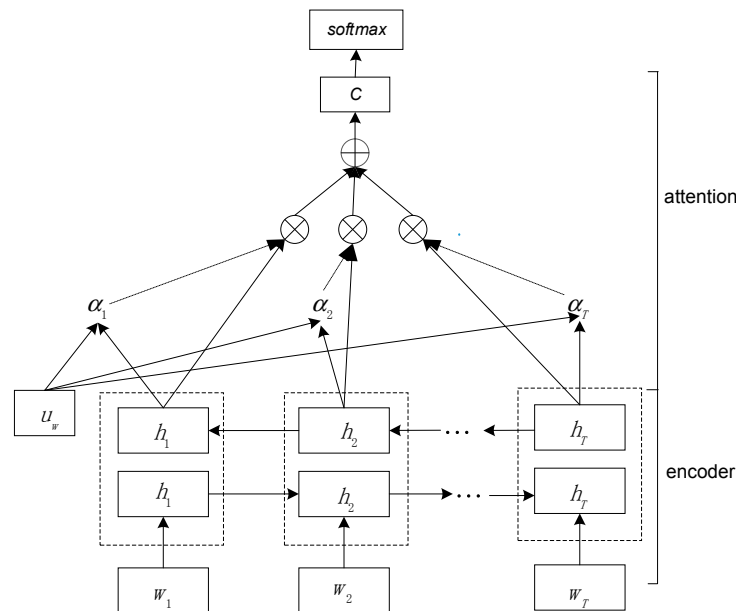


Figure 3. Workflow of BiGRU neural network model with attention mechanism.

#### 3.2.1. BiGRU

In our model, BiGRU, a bi-directional recurrent neural network model, is used to map a sequence of word vectors of the document to sentiment categories. In BiGRU, a gated recurrent unit (GRU) uses gates to resolve the problem of gradient vanishing for preserving the long-distance information. A GRU has two gates: a reset gate and an update gate, as illustrated in Figure 4. The reset gate  $r_t$  controls how past information contributes to the candidate state  $\vec{h}_t$ ; the update gate  $z_t$  determines how past information is preserved and how new information is added.

At time  $t$ , we compute the hidden vector  $\vec{h}_t$  of the forward GRU:

$$\vec{h}_t = \begin{cases} (1 - \vec{z}_t) \odot \vec{h}_{t-1} + \vec{z}_t \odot \vec{h}_t, & \text{if } t > 0 \\ 0, & \text{if } t = 0 \end{cases}; \quad (7)$$

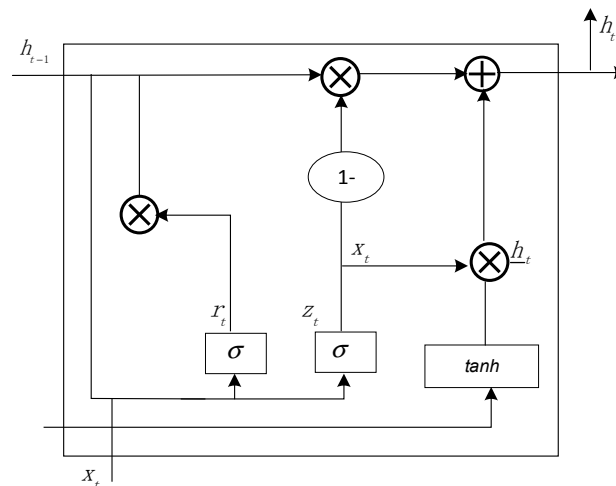
$$\vec{h}_t = \tanh\left(\vec{W}x_t + \vec{U}\left[r_t \odot \vec{h}_{t-1}\right]\right); \quad (8)$$

$$\vec{r}_t = \sigma(\vec{W}_r x_t + \vec{U}_r \vec{h}_{t-1}); \quad (9)$$

$$\vec{z}_t = \sigma(\vec{W}_z x_t + \vec{U}_z \vec{h}_{t-1}). \quad (10)$$

Similarly, we compute the hidden vector  $\overleftarrow{h}_t$  of the backward GRU. Then, the hidden vector  $h_t$  of BiGRU is calculated as follows:

$$h_t = \begin{bmatrix} \vec{h}_t \\ \overleftarrow{h}_t \end{bmatrix}. \quad (11)$$



**Figure 4.** Structure diagram of the gated recurrent unit (GRU).

### 3.2.2. Attention Mechanism

Since not all words contribute equally to the meaning of the text, we added the attention mechanism to the BiGRU neural network model to emphasize the words important to the meaning of the text during sentiment classification. Then, a document vector is formed by these word vectors weighed by their importance on the document, and then, the document is finally classified.

The importance of a word in a document can be computed by the context of the word. The BiGRU model uses information in both forward and backward directions to get the contextual information, which captures the word connotation. For a given text  $c$ , it contains  $T$  words;  $w_t$  denotes the  $t$ -th word in a document and  $x_t$  denotes the  $t$ -th word vector,  $t \in [1, T]$ . Forward GRU reads the text  $c$  from  $w_1$  to  $w_T$ , while the backward GRU does it in reverse. Specifically, we firstly use a one-layer multilayer perceptron (MLP) to get  $u_t$  as the hidden representation of word annotation  $h_t$ , and then, we use a word-level context vector  $u_w$  which is randomly initialized to measure the importance of the word as the similarity of  $u_t$ . The context vector  $u_w$  can be seen as a high-level representation of the informative word and the value of  $u_w$  is updated during the training process. Finally, we get the importance weight  $\alpha_t$  normalized by softmax function.

$$u_t = \tanh(W_w h_t + b_w); \quad (12)$$

$$\alpha_t = \frac{\exp(u_t^\top u_w)}{\sum_t \exp(u_t^\top u_w)}. \quad (13)$$

After that, we compute the document vector as a weighted sum of the word annotations at the current time in the decoded state using Equation (14).

$$c = \sum_t \alpha_t h_t. \quad (14)$$

### 3.2.3. Document Classification

As shown in Figure 3, the vector  $c$ , calculated as the feature representation of the document, is used to calculate the probability that the document belongs to each category:

$$p = \text{softmax}(W_c c + b_c). \quad (15)$$

We compute the negative log likelihood of the correct labels as training loss:

$$L = - \sum_d \log p_d y, \quad (16)$$

where  $y$  is the label of document  $d$ .

## 4. Experiment Setting

To evaluate the performance of our BiGRULA model for sentiment classification, we first tested how the lda2vec, the topic-enhanced word-vector-encoding method used in our model, performed in document classification compared to other text-encoding methods. We then applied the BiGRULA model to the well-known IMDB movie review dataset.

### 4.1. Evaluation of Lda2vec

In order to evaluate the lda2vec model, we considered using Liblinear [21] classifier, a special type of support vector machine, to permit multiclass classification on the 20 NewsGroup dataset from Scikit-learn, which has approximately 20,000 newsgroup reviews. Liblinear is a linear classifier which can train large-scale data in a small amount of time. We used it to evaluate the performance of word vectors trained by lda2vec and compared it with other word embedding. The reason we chose Liblinear instead of other popular classifiers such as RandomForest is that we wanted to compare our results with those from literature [13], where this algorithm was used. Common models of text encoding include BOW (bag-of-words), skip-gram, PV [22] (paragraph vector), LDA, glove, and so on. Among them, the PV model includes PV-DM (distributed memory model of paragraph vectors) and PV-DBOW (distributed bag of words version of paragraph vector).

In the lda2vec model, we used pretrained word vectors (GoogleNews-vectors-negative300.bin, which includes a vocabulary of three million words and phrases trained from a Google News dataset) to train the topic-enhanced word vectors and document vectors in the 20 NewsGroup train set, as shown in Figure 2. Table 1 shows the results of text classification of the 20 NewsGroup test set. We can observe that lda2vec outperformed all baselines. This indicates that lda2vec can capture richer and more precise information of the documents.

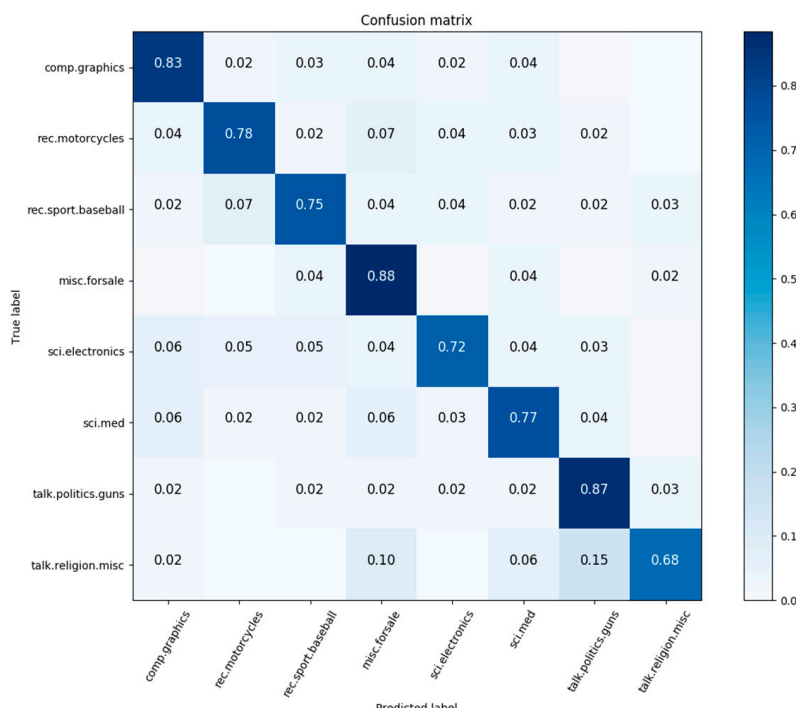
**Table 1.** Performance comparison of lda2vec with other text encoding methods.

Model	Accuracy	Precision	Recall	F-Measure
BOW	79.2%	79.5%	79.0%	79.0%
LDA	72.2%	70.8%	70.7%	70.0%
Skip-Gram	75.4%	75.1%	74.3%	74.2%
PV-DM	72.4%	72.1%	71.5%	71.5%
PV-DBOW	75.4%	74.9%	74.3%	74.3%
Glove	76.6%	76.7%	76.5%	76.5%
<b>lda2vec</b>	<b>80.2%</b>	<b>81.1%</b>	<b>80.0%</b>	<b>80.0%</b>

BOW (Bag-of-Words), PV-DM (Distributed Memory Model of Paragraph Vectors), PV-DBOW (Distributed Bag of Words Version of Paragraph Vector), Glove (Global Vectors for Word Representation).



To further illustrate the performance of lda2vec, Figure 5 shows the confusion matrix with the percentages of samples for each class predicted by Liblinear. Each column of the confusion matrix represents the predicted label (output class), while each row represents the true label (target class). As shown in Figure 5, the largest percentage of true positive classification was 88% for category misc.forsale. The smallest percentage of true positive classification was 68% for talk.religion.misc, while 15% of talk.religion.misc was misclassified as talk.politics.guns.



**Figure 5.** Confusion matrix for multi-class classification with lda2vec word embedding.

## 4.2. Evaluation of BiGRULA Model

### 4.2.1. Dataset and Parameter Settings

We used the IMDB movie review dataset [23] from Scikit-learn to evaluate our sentiment classification model. The dataset had 50,000 reviews, allowing no more than 30 reviews per movie. The whole dataset was split into 25,000 training samples and 25,000 testing samples. There were 25,000 positive reviews and 25,000 negative reviews in this dataset.

In our model, we used Google News-vectors-negative 300.bin as the pretrained word vectors used in the lda2vec module of BiGRULA. The number of topics for each document was set to 10 after evaluating its value from four to 20, where 10 was identified as the best. We consider the 10,000 top most frequent words in IMDB, and set the sequence length as 250. In the training of BiGRU with attention, we set the hidden size as 150, the attention size as 50, the batch size as 256, and the keep probability of training samples as 0.8.

### 4.2.2. Results and Analysis

To evaluate our BiGRULA model, we compared our results with those of other various neural networks. At the same time, we also trained a set of BiGRU models with attention mechanism using three other common word-vector-encoding methods to evaluate the performance of lda2vec. These models were all trained on the IMDB dataset with an equal number of parameter settings, and the results are presented in the Table 2.

**Table 2.** Comparison of models for sentiment classification.

Model		Train set		Test set	
		Loss	Accuracy	Loss	Accuracy
BiGRU with attention	Word Embedding				
	Random	0.177	0.944	0.357	0.864
	Skip-gram	0.179	0.927	0.314	0.869
	Glove	0.170	0.945	0.375	0.872
	Lda2vec	0.211	0.914	0.259	0.894
CNN	0.230	0.907	0.287	0.881	
LSTM	0.011	0.997	1.094	0.812	
CNN + LSTM	0.198	0.925	0.346	0.858	
G_TF-IDF + FPCD + NB [24]	–	–	–	0.870	
Word2vec + KNN [24]	–	–	–	0.773	
FPCD + SVM [24]	–	–	–	0.857	
N-gram + SVM [6]	–	–	–	0.889	
N-gram + NB [6]	–	–	–	0.862	
N-gram + ME [6]	–	–	–	0.885	
Word2vec + LR [25]	–	–	–	0.847	

CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), CNN + LSTM (CNN is used as a feature extractor, and LSTM is used as a classifier), G\_TF-IDF+FPCD (FPCD feature vectors combined with the generalized TF-IDF vectors)+NB (Naïve Bayes), Word2vec + KNN (K-Nearest Neighbor), FPCD (a frequent, pseudo-consecutive phrase feature with high discriminative ability) + SVM (Support Vector Machine), N-gram+SVM (Support Vector Machine n-gram classifier), N-gram + NB (Naïve Bayes n-gram classifier), N-gram + ME (Maximum Entropy n-gram classifier), Word2vec + LR (Logistic Regression).

We found that the topic-enhanced word vectors learned from lda2vec achieved the lowest accuracy of 0.914 compared to three other word-embedding methods over the training set. However, it achieved the highest accuracy of 0.894 on the test dataset compared to 0.864, 0.869, and 0.872 of the other three word-embedding methods. When comparing the BiGRU RNN network with other neural networks such as Convolutional Neural Network (CNN), Long Short-term Memory (LSTM), and CNN+LSTM, our BiGRULA achieved better accuracy with 0.894 over the test dataset than the other three network models with accuracies of 0.881, 0.812, and 0.858, respectively.

Table 2 also shows the comparison accuracy values with other known machine learning approaches as available in literature [6,24,25] using the IMDB dataset. FPCD feature vectors combined with the generalized TF\_IDF vectors + Naïve Bayes (G\_TF-IDF + FPCD + NB), Word2vec + K-Nearest Neighbor (Word2vec + KNN), and frequent, pseudo-consecutive phrase feature with high discriminative ability + Support Vector Machine (FPCD + SVM) achieved the highest accuracy among their feature extraction methods, while, compared to our model, their accuracy values still could not compare. When we compared our model with Support Vector Machine (SVM), Naïve Bayes (NB), and Maximum Entropy (ME), which had the best accuracy values among *n*-gram methods, and with word2vec + LR, which had the best value among three different features, our model still showed the best accuracy value.

## 5. Application of BiGRULA to Sentiment Analysis of Tourism Reviews

Here, we demonstrate the utility of our BiGRULA model in tourism review analysis. We chose the ChnSentiCorp-Htl-unba-10000 hotel review dataset, a set of Chinese hotel reviews collected by Songbo Tan from Ctrip [26]. It had 7000 positive reviews and 3000 negative reviews. Through our model, we extracted and utilized the useful information hidden in the hotel review data and acquired customers' sentimental attitude toward the hotels.

Firstly, we used lda2vec to extract the topics of the hotel reviews. We used sgns.Weibo.word, a set of pretrained Chinese word vectors [27]. Common topics of hotel reviews include aspects about

hotel environment, hygiene, transportation, diet, supporting facilities, price, hotel service, tourism network service, entertainment, and surroundings. Therefore, we set the number of topics as 10 for the BiGRULA model in our experiments. By training the model of lda2vec, we acquired the topic-enhanced word vectors and the topics of these hotel reviews after 400 epochs. Furthermore, we estimated the most relevant terms within the selected topic [28].

$$relevance(w|T) = \lambda * p(w|t) + (1 - \lambda) \times \frac{p(w|t)}{p(w)}, \quad (17)$$

where  $p(w)$  indicates the probability of term  $w$ ,  $p(w|t)$  indicates the probability of term  $w$  under topic  $t$ , and  $\lambda$  determines the weight given to  $p(w|t)$ .

The topics and their most relevant terms extracted by lda2vec are displayed in Table 3. As the baseline, we also trained an LDA model with 50,000 epochs on the same hotel review dataset. The number of topics was also set as 10, and the results are present in Table 3. From the results in Table 3, we can observe that the topics extracted by LDA and lda2vec were all incoherent. We suspect that the main reason was that these hotel reviews were much shorter compared to other types of documents that LDAs are commonly applied to. This makes the models unable to extract the topics very well. However, with close examination, we found that the topics extracted by lda2vec were closer to the common topics of the hotel reviews as recognized by humans when compared to LDA. These results demonstrate that lda2vec performed better than LDA in topic extraction.

**Table 3.** Comparison of topic terms extracted by LDA and lda2vec.

	LDA	lda2vec
topic 0	careful; Shandong; you; sleeps; takes; counter; off-season; Sanya; usually	hotel; service; environment; transportation; taxi; opposite; shopping; credit card
topic 1	you; experience; rotate; for a while; reserved rights; cold; enough; switch	room; bathroom; too; carpet; facility; stale; small; poor
topic 2	you; counter; rotate; carefully; reserved; for a while; rights; recruitment; talent	service; Comfortable; enthusiasm; hotel; feel; attentive; features; tell
topic 3	you; bank; careful; signing; settlement	check-in; service; hotel; waiter; room; front desk; warm; guest
topic 4	calling; most; expensive; ladies; items; thank you very much; direction	front desk; ask; phone; tell; Ctrip; waiter; call
topic 5	you; reserved; rotate; rights; careers; copyright1999; agent; advertising business; experience	room; nice; large; bathroom; bed; feeling; facilities; comfortable
topic 6	too few; rotations; enough; counters; most; settlements; a while; banks; off-season; usually	room; check-in; feel; facilities; clean; comfortable; disadvantages; will
topic 7	Careful; slightly; picking up; Shandong; Zhengzhou; expensive; rotating; enough; one bottle; climb	breakfast; hotel; eat; restaurant; variety; taxi; price; delicious
topic 8	you; Shandong; you; apologize; experience; careful; related; Square meters; bank; thank you very much.	hotel; transportation; downtown; price; airport; service; next time
topic 9	you; settlement; counter; a little bit; make; front, climbing; direction; picking up	sound; room; night; soundproofing; hygienic; air conditioning; bathroom; windows

In addition, to illustrate how informative the extracted terms by lda2vec were, we computed the term saliency for each term and got the top most salient terms [29].

$$saliency = P(w) \times distinctiveness(w), \quad (18)$$

where  $w$  denotes a given term, and we define the distinctiveness of  $w$  as the Kullback–Leibler divergence between  $P(T|w)$  and  $P(T)$ :

$$\text{distinctiveness}(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}, \quad (19)$$

where  $P(T|w)$  is the conditional probability that term  $w$  belongs to topic  $T$ .

This calculated distinctiveness describes how informative the specific term  $w$  is for determining the generating topic. If a term occurs in all topics, which tells us little about the document's topical mixture, the term would receive a low distinctiveness score.

Ranking all the terms by their term saliency, we got the top-30 most salient terms, as shown in Figure 6. We can observe that the most salient terms were positive and it showed that the customers' general impression about the hotel was good. The top most salient terms such as "environment", "price", "shopping", "transportation", and so on, tell us some topical information. From Figure 6, we can find that the frequent terms were not always salient. For example, the frequency of "good" was ranked as first, while its saliency was far below first. In addition, we can observe that most top salient terms can be found in topic terms extracted by lda2vec, as shown in Table 3, which demonstrated the effectiveness of the lda2vec model.

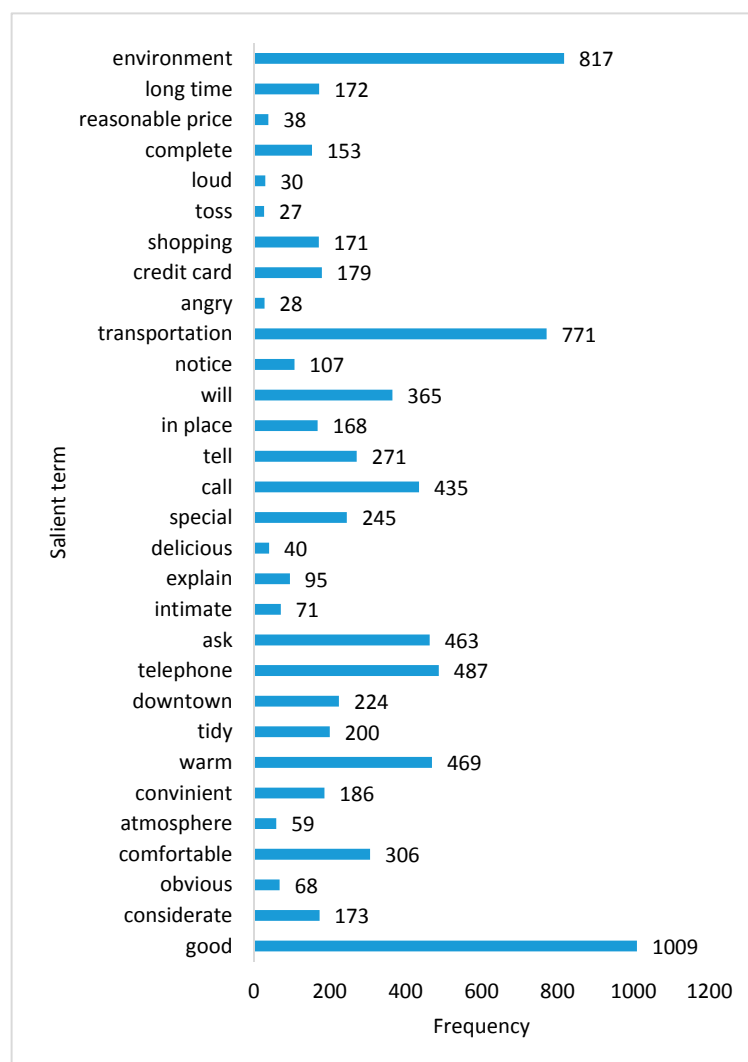
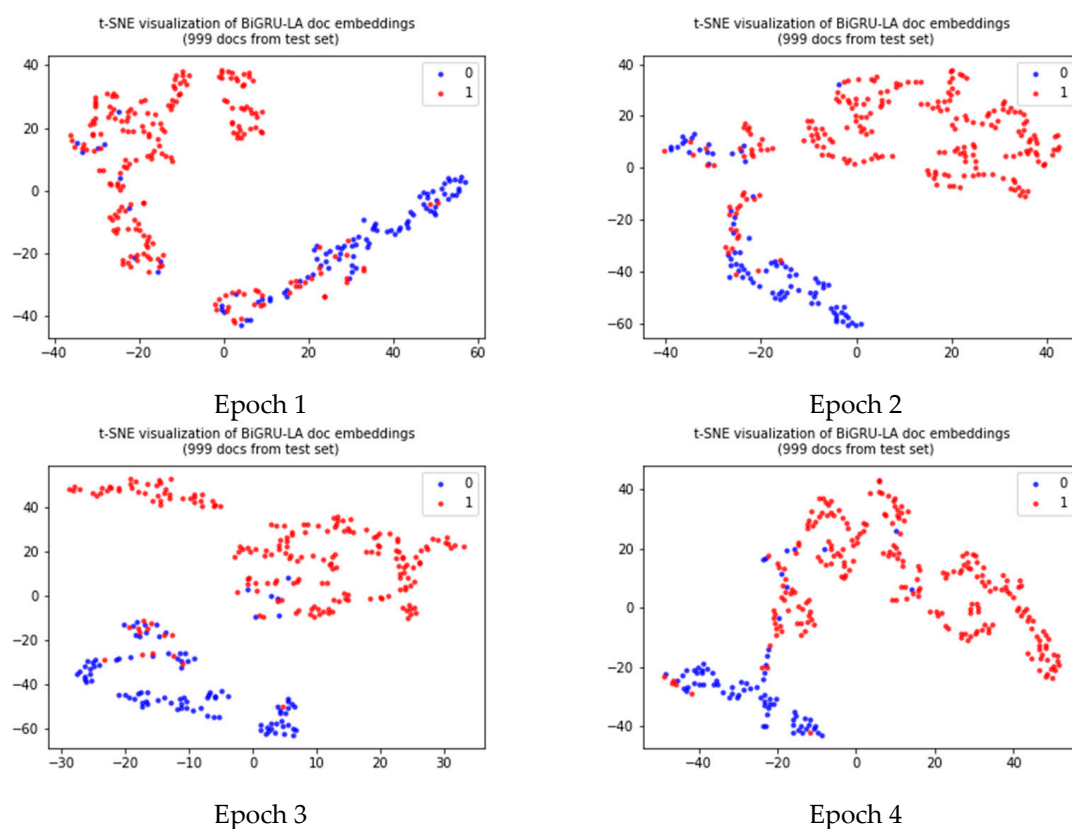


Figure 6. Top-30 most salient terms from top to bottom.

Secondly, we input the pre-trained word vectors trained by *lda2vec* into the BiGRU model with the attention mechanism for sentimental analysis of hotel reviews. In our experiment, we set the sequence length and the hidden size as 100, the attention size as 50, the batch size as 300, and the keep probability of training samples as 0.8. After training four epochs, our model converged to high classification accuracy of 93.1% in our binary sentiment classification over the hotel reviews.

In order to further observe the process of sentimental classification and demonstrate the effectiveness of our model, we used t-distributed stochastic neighbor embedding (t-SNE), a nonlinear dimensionality reduction algorithm [30] to visualize the results. Figure 7 demonstrates the t-SNE visualization of BiGRULA document embedding of four epochs from the test set. We can see that the negative reviews (tagged 0) and positive reviews (tagged 1) were gradually separated through four epochs. This further demonstrates that our model can differentiate the negative and positive emotions within the review comments well.



**Figure 7.** Evolution of the separation of positive and negative reviews during the training process.

## 6. Conclusions and Future Work

In this work, we proposed the BiGRULA model for sentimental classification and applied it to Chinese hotel review analysis. This model is characterized by its synergistic combination of *lda2vec*, a topic-enhanced word-embedding approach, with a bidirectional recurrent neural network model with attention mechanism. Through experimental evaluation, we showed that this model can achieve better performance than other popular neural network models such as CNN, LSTM, and CNN+LSTM. Application of our BiGRULA model to hotel review analysis showed that it can extract rich information from the text dataset, which is also closer to the meaning of the text. These features extracted from the text further improve the performance of succeeding sentimental classification.

Several aspects of our BiGRULA model can be further improved. We note that BiGRULA model only has a word-level attention mechanism, which may limit the training ability of the model. This issue may be addressed by the hierarchical attention network (HAN) as proposed in Yang [31],

which builds the representation of a sentence using a word-level attention mechanism and builds a document classifier using a sentence-level attention mechanism. In the near future, we aim to add the sentence-level attention into our model.

Our study has some practical implications and applications. It can extract topics from online hotel reviews, which can give hoteliers insight into these reviews and can capture different determinants of guest satisfaction, which allows them to realign their strategies in service and product development, such as meaningful hotel competitive sets to better reflect guests' perspective. Also, our model BiGRULA can analyze consumers' sentiment and satisfaction, which can effectively help hoteliers evaluate the performance of the hotel operation and further formulate their strategies in the marketplace. Our model proved effective in the Chinese hotel consumer online review sentiment prediction. Our algorithm can also be used by hotel recommendation websites or booking platforms such as TripAdvisor, Ctrip, and so on to automatically rank hotels by sentiment analysis of their online reviews.

**Author Contributions:** Q.L., S.L., and J.J.H. conceived of and designed the study. Q.L. and J.H. worked on the algorithm design. S.Z. implemented the baseline methods. Q.L. and J.J.H. wrote the manuscript, made the figures, and reformatted the manuscript. Q.L. and J.J.H. revised and polished the manuscript. All authors read and approved the final manuscript.

**Funding:** This work is supported by the National Natural Science Foundation of China under Grant Nos. 91746116 and 51741101, National Science and Technology Supporting Plan (2014BAH05F01, 2014BAH05F02, 2014BAH05F03), and the Science and Technology Project of Guizhou Province under Grant Nos. JZ[2014]2001, [2014]6021, Talents [2015]4011, and Collaborative Innovation [2015]02.

**Acknowledgments:** We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. We also would like to thank Yong Yao for his help in programming.

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

## References

1. Zheng, X.; Luo, Y.; Sun, L.; Zhang, J.; Chen, F. A Tourism Destination Recommender System Using Users' Sentiment and Temporal Dynamics. *J. Intell. Inform. Syst.* **2018**, *6*, 1–22. [CrossRef]
2. Ren, G.; Hong, T. Investigating Online Destination Images Using a Topic-Based Sentiment Analysis Approach. *Sustainability* **2017**, *9*, 1765. [CrossRef]
3. Li, Q.; Wu, Y.; Wang, S.; Lin, M.; Feng, M.; Wang, H. VisTravel: Visualizing Tourism Network Opinion from the User Generated Content. *J. Visual.* **2016**, *19*, 489–502. [CrossRef]
4. Serna, A.; Gerrikagoitia, J.K.; Bernabe, U.; Ruiz, T. A Method to Assess Sustainable Mobility for Sustainable Tourism: The Case of the Public Bike Systems. In *Information and Communication Technologies in Tourism 2017*; Schegg, R., Stangl, B., Eds.; Springer: Berlin, Germany, 2017.
5. Application of Social Media Analytics: A Case of Analyzing Online Hotel Reviews. Available online: <https://www.emeraldinsight.com/doi/abs/10.1108/OIR-07-2016-0201> (accessed on 4 September 2018).
6. Tripathy, A.; Agrawal, A.; Rath, S.K. Classification of Sentiment Reviews Using N-Gram Machine Learning Approach. *Expert Syst. Appl.* **2016**, *15*, 117–126. [CrossRef]
7. Hu, Y.; Li, W. Document Sentiment Classification by Exploring Description Model of Topical Terms. *Comput. Speech Lang.* **2011**, *25*, 386–403. [CrossRef]
8. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. *A Convolutional Neural Network for Modelling Sentences*; Cornell University Library: New York, NY, USA, 2014.
9. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent Convolutional Neural Networks for Text Classification. In Proceedings of the National Conference on Artificial Intelligence, Austin, TX, USA, 25–29 January 2015. Available online: <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/download/9745/9552> (accessed on 4 September 2018).
10. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. Available online: <https://arxiv.org/pdf/1301.3781.pdf> (accessed on 4 September 2018).
11. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.



12. Wang, Z.; Ma, L.; Zhang, Y. A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec. In Proceedings of the IEEE First International Conference on Data Science in Cyberspace, Changsha, China, 13–16 June 2016.
13. Liu, Y.; Liu, Z.; Chua, T.; Sun, M. Topical Word Embeddings. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–29 January 2015. Available online: <https://pdfs.semanticscholar.org/9a0a/f9e48aad89512ce3e24b6a1853ed3d5d9142.pdf> (accessed on 4 September 2018).
14. Yao, L.; Zhang, Y.; Chen, Q.; Qian, H.; Wei, B.; Hu, Z. Mining Coherent Topics in Documents Using Word Embeddings and Large-Scale Text Data. *Eng. Appl. Artif. Intell.* **2017**, *64*, 432–439. [CrossRef]
15. Zhang, D.; Luo, T.; Wang, D. Learning from LDA Using Deep Neural Networks. In *Natural Language Understanding and Intelligent Applications*; Lin, C.Y., Xue, N., Zhao, D., Huang, X., Feng, Y., Eds.; ICCPOL 2016, NLPCC 2016, Lecture Notes in Computer Science, vol. 10102; Springer: Berlin, Germany, 2016.
16. Moody, C.E. Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec. Available online: <https://www.datacamp.com/community/tutorials/lda2vec-topic-model> (accessed on 4 September 2018).
17. Dieng, A.B.; Wang, C.; Gao, J.; Paisley, J. TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency. Available online: <https://arxiv.org/abs/1611.01702> (accessed on 4 September 2018).
18. Li, S.; Zhang, Y.; Pan, R.; Mao, M.; Yang, Y. Recurrent Attentional Topic Model. In Proceedings of the National Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017. Available online: <http://www.shuangyin.li/publications/ratm.pdf> (accessed on 4 September 2018).
19. Raffel, C.; Ellis, D.P.W. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. 2015. Available online: <https://arxiv.org/abs/1512.08756> (accessed on 4 September 2018).
20. Wang, L.; Yao, J.; Tao, Y.; Zhong, L.; Liu, W.; Du, Q. A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization. Available online: <https://arxiv.org/abs/1805.03616> (accessed on 4 September 2018).
21. Fan, R.; Chang, K.; Hsieh, C.; Wang, X.; Lin, C. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
22. Le, Q.V.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014. Available online: <http://proceedings.mlr.press/v32/le14.pdf> (accessed on 4 September 2018).
23. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011.
24. Chen, X.; Xue, Y.; Zhao, H.; Liu, X.; Hu, X.; Ma, Z. A Novel Feature Extraction Methodology for Sentiment Analysis of Product Reviews. *Neural Comput. Appl.* **2018**, *4*, 1–18. [CrossRef]
25. Zhu, J.; Yu, W. Binary Sentiment Analysis on IMDb Movie Reviews. Available online: <https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a055.pdf> (accessed on 4 September 2018).
26. Tan, S. ChnSentiCorp-Htl-unba-10000. Available online: [https://download.csdn.net/download/sinat\\_30045277/9862005](https://download.csdn.net/download/sinat_30045277/9862005) (accessed on 5 September 2018).
27. Li, S. sgns.Weibo.word. Available online: <http://github.com/Embedding/Chinese-Word-Vectors> (accessed on 5 September 2018).
28. Sievert, C.; Shirley, K.E. LDAvis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Baltimore, MD, USA, 27 June 2014.
29. Chuang, J.; Manning, C.D.; Heer, J. Termite: Visualization Techniques for Assessing Textual Topic Models. in *Advanced Visual Interfaces*. Available online: <http://vis.stanford.edu/files/2012-Termite-AVI.pdf> (accessed on 4 September 2018).
30. Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Available online: <http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf> (accessed on 4 September 2018).
31. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016.

