Reputation Analysis Based on Weakly-Supervised Bi-LSTM-Attention Network

XIANG, KUN

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学研究科編
(巻 / Volume)
63
(開始ページ / Start Page)
1
(終了ページ / End Page)
12
(発行年 / Year)
2022-03-24
(URL)
https://doi.org/10.15002/00025364

Reputation Analysis Based on Weakly-Supervised Bi-LSTM-Attention Network

19R8107 XIANG KUN
M2 IIST
Applied Informatics, Graduate School of Science and Engineering
Hosei University
Supervisor: Professor Akihiro Fujii

***Abstract* — Due to the emergence of rapid, mass-produced information in the Web2.0 era, a large amount of weakly labeled information (star ratings, etc.) has been widespread. The Weakly-Supervised Deep Embedding (WDE) model is a good choice for utilizing this kind of data. The ratings are treated as weakly-supervised signals for pre-training, fine tuning the whole model with a small amount of manually labeled data. In this research, we proposed to change the original unidirectional transmission into bidirectional in the LSTM layer to capture the semantics in both directions, and an attention mechanism is introduced, which is helpful to capture the important information in the context and improve the accuracy of sentiment classification. Finally, we use TF-IDF and LDA topic models to mine the review topics and extract the consumers' opinions on different sentiment polarities.**

Keywords: Sentiment Classification, Deep Learning, Opinion Mining, Weak Supervision.

1 INTRODUCTION

Product reviews have an invisible impact on consumer's judgement and affect their purchasing decisions with the rapid popularization of user-generated content (UGC) in the Web2.0 era. Among them, tourism products play a special role because of their unique product properties [4]. Abundant species and high value and complexity of obtaining a service process makes consumers more vigilant for making choices. To this end, tourist reviews on various websites have become an important reference for tourists, have impact on people's judgment of travel, and then affect consumers' purchasing decisions [2].

Recently, wildlife tourism is more and more rapid and extensive in development. According to statistics, wildlife tourism accounts for 20-40% of the global tourism industry. Every year, millions of tourists around the world participate in wildlife recreational tourism activities. According to the report of the World Tourism Organization (WTO) in 2019, wildlife tourism contributes USD 120.1 billion to the global GDP, five times that of illegal trade (USD 23 billion). At present, 7% of the world's tourism is related to wildlife tourism, with an annual growth rate of 3%, which is a huge market. This leads to the ethical thinking between the healthy development of wildlife tourism and wildlife protection. From last year, with the outbreak of COVID-19, some raise the opinion that wildlife is the source or host of various infectious diseases. The relationship between wildlife and humans is at an important turning point. People's attitude towards wildlife tourism is becoming a hot issue. In this research, we take wildlife tourism reviews on TripAdvisor as a training dataset.

From the necessity of text mining, in today's network information explosion, it is very important to mine effective topics from the massive amount of information [32]. From the technical means of sentiment analysis, the existing classification methods are mainly based on the following type of methods: sentiment word dictionary and machine learning. In the traditional sentiment classification methods, analysis is usually carried out by constructing a specific dictionary in a certain field, or by using machine learning methods relying on feature

selection. Such classification methods are deeply affected by professional knowledge, experience and complete domain-related knowledge systems, and it has unsatisfactory performance in other fields. What's more, when it involves a foreign language, it may be very difficult. Recently, deep learning has emerged as an effective means for solving sentiment classification problems [7]. However, it needs a large-scale of effectively labeled datasets to help training model parameters.

Due to the emergence of rapid, mass-produced information, a large amount of weakly labeled information (star ratings, etc.) has been widespread. An example is shown in figure 1. There is a certain correlation between this kind of information and sentiment orientation of comments, but there is also some noise [3][4][5].

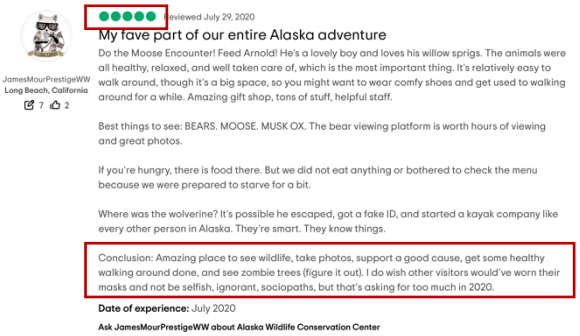

Fig. 1. A 5-star review with negative words (noise)

The WDE model [5] is a good choice for utilizing of weakly labeled data, the ratings are treated as the weakly-supervised signals for pre-training and fine-tune the whole model with the small-scale of manually labeled data. The results show that the WDE-LSTM model performs well in classification accuracy and small-scale training datasets. Based on WDE-LSTM, in this work, we proposed to change the original unidirectional transmission into bidirectional in the LSTM layer to capture the semantics in both directions, and an attention mechanism is introduced, which is helpful to capture the important information in the text and improve the accuracy of sentiment classification [3][5]. The enhanced model is named **WDE-BiLSTM-Attention**. We can find notable improvement of classification accuracy and efficiency in the result. Finally, we use TF-IDF and LDA topic models to mine the review topics, and extract the consumers' opinions on different sentiment polarities.

To sum up, the innovations of this research are:

- Propose an enhanced method based on the WDE-LSTM model, change the original unidirectional transmission into bidirectional in the LSTM layer, and introduce an attention mechanism in order to capture important information.
- Make full use of a large-scale of weakly labeled information on the Internet, and train the model with a small-scale sample dataset, which greatly improves the accuracy and efficiency.
- COVID-19 is widely affecting people's daily lives, the relationship between wildlife and human beings is on a turning point. We focus on the ecological ethics contradiction of wildlife tourism, from the perspective of tourists, and extract their opinions on wildlife. Understanding the interaction between wildlife and tourists is of great significance for the sustainable development of wildlife tourism industry and ecological ethics.

The rest of this paper is organized as follows: Section 2 is related works, and Section 3 introduces the basic model WDE and our enhancement. We also explain the construction process in detail. Section 4 introduced our experiments and results in improvement. What's more, we mine topics of our dataset, and find some interesting conclusions which have practical significance nowadays. Finally, we conclude our work and discuss about the deficiencies and what we can do in future work.

2 RELATED WORKS

Sentiment analysis (SA) is also known as tendency analysis and opinion mining. It is quite a long-standing research topic [9][14] in natural language processing (NLP). The rapid development of Internet technology has brought new means of information transmission through UGC, in the form of blogs, social media, website reviews, e-commerce website, etc. [1][2]. Generally speaking, the goal of sentiment analysis is to clarify the reviewers' attitude at the document level, sentence level and aspect level [3][12]. Sentiment analysis usually relies on two types of techniques, i.e., lexicon based [13][24] and machine learning based [18] techniques. The most basic task of sentiment analysis is identifying the polarity and subjectivity of documents using a combination of machine learning, information retrieval, and natural language processing techniques [42][43], to divide a given comment text into three categories: positive, negative and neutral at different levels. On this basis, we can also set the goal of multi-polar emotion classification according to the actual problems, such as dividing news comments into "sadness," "optimism" and "anger" [27]. However, in this research, we just consider two categories of positive and negative.

2.1 Machine Learning for Sentiment Analysis

B. Pang *et al.* [14][21] put forward a standard machine learning method for the first time. They studied the sentiment analysis of movie reviews in the form of text on the Internet and converted the problem of text polarity classification into solving the minimum segmentation problem of a sentence connection graph to classify the text sentiment. Sanjiv Ranjan Das *et al.* applied the theory of sentiment analysis to the emotion tendency of foreign stock websites through the research of sentiment analysis [16]. Wiebe *et al.* studied the classification of customers' sentiment from the perspective of subjective sentence and objective sentence classification and proposed the classification method. Dave *et al.* tried the effect of different types of feature combinations in emotion classification and compared the effect of the unigrams feature with the bigrams feature in emotion classification. The experimental results show that the classification accuracy of the bigrams feature is higher than that of the

unigrams feature under the same conditions [22]. Li *et al.* proposed a novel approach named eVector to construct an emotion lexicon and to generate a new feature representation of text. They compared the system with supervised machine learning classifier(SVM) based on bag-of-words(BoW) and present the system for the Sina Weibo texts on both the document and sentence level. The result showed that both systems can classify emotion significantly better than random guess [19]. Fei *et al.* proposed to use sentence phrase patterns to classify the emotional tendency of the text which the score is calculated by constructing each sentence phrase pattern in the text [17]. Medhat *et al.* used Separate Bayes, Maximum Entropy model and SVM three machine learning methods to classify the sentiment of the corpus composed of film reviews [20].

The machine learning-based classification method focuses on the research of feature engineering and the quality of the annotation dataset [6]. This kind of method still relies on artificial design and is easily affected by human factors in the research process [8]. Moreover, it has poor promotion ability in different fields. It is not necessarily applicable to other fields. The performance of most classification models also depends on the quality of the annotation dataset while obtaining high quality annotation datasets requires a lot of labor costs [36][37].

2.2 Deep Learning for Sentiment Analysis

Since the introduction of unsupervised training in 2006, deep learning has gradually become a hot research direction [15]. With the rapid development of deep learning, the task of sentiment classification is realized [25].

Socher *et al.* [48] proposed a series of Recursive Neural Network (RNN) models for sentiment classification. In addition, Convolutional Neural Network (CNN) models have also been applied in the field of sentiment analysis in recent years [33][34]. Santos *et al.* proposed CNN for sentiment classification of short texts and achieved remarkable results [27]-[31]. Zhu *et al.* proposed Long Short-Term Memory (LSTM) to separate comment statements into word sequences to solve the problem of sentiment classification [29][31]. LSTM can capture the long-term dependencies in the comments and understand the emotional semantics of comments as a whole [10][29]. Deep neural networks adopt a bionics principle to imitate the hierarchical structure of the human brain and have the expression ability of exponential times that of shallow computing models [45][46].

Embedding Learning. In the text-understanding tasks, more and more attention has been paid to learning distributed text embedding using a neural network model [11], and emotion classification is a popular method. Farman Ali *et al.* proposed an ontology and Latent Dirichlet Allocation (OLDA)-based topic modeling and word embedding approach for sentiment classification which retrieves transportation content from social networks, removes irrelevant content to extract meaningful information and generates topics and features from extracted data [23][47]. Le and Mikolov developed an unsupervised embedded learning method for sentences, paragraphs and documents [7][33]. Then, two simple network models are proposed. Kiros *et al.* proposed an unsupervised model, which extended a skip-gram model to the sentence level [35]. However, few works tried to use review ratings to train deep sentiment classifiers for sentences. Wei Zhao *et al.* firstly proposed the WDE framework to make use of rating information as weak labels for training deep sentence sentiment classifiers [5]. It attempts to construct a weakly-supervised deep learning framework and effectively exploit weakly labeled datasets.

Attention Mechanism. Inspired by the recent success of attention-based neural networks [37], Tang *et al.* introduced a deep memory network for aspect level sentiment classification which is fast and effective by leveraging both content and location information, and learning better context weight and text representation [37]. Cheng *et al.* proposed a Hierarchical Attention (HEAT) Network for aspect-level sentiment classification which contains a Hierarchical Attention module, consisting of aspect attention and sentiment attention [39]. Wang and Chen proposed Dependency-Attention-based Long Short-Term Memory Network (DAT-LSTM) and Segmented Dependency-Attention-based Long Short-Term Memory Network (Seg-DAT-LSTM) for target-dependent sentiment analysis. The dependency-attention mechanism utilizes dependency relation to fully capture long-range information for a certain target [40]. Chen *et al.* proposed a novel framework based on neural networks to identify the sentiment of opinion targets in a comment review which adopts a multiple-attention mechanism to capture sentiment features separated by a long distance, so that it is more robust against irrelevant information [38]. Ma *et al.* proposed the interactive attention networks (IAN) to interactively learn attentions in the contexts and targets and generate the representations for targets and contexts separately which can well represent a target and its collocative context for better performance on sentiment classification [41].

3 MODEL IMPLEMENT

3.1 The Classic WDE Network Architecture

Wei Zhao *et al.* [5] proposed to learn an embedding space which can properly reflect data's semantic distribution in the weakly-supervised training phase. The final classification layer is added in the following supervised training phase, in order to learn the final prediction model.

Embedding Training with Ratings. The review sentences can be divided into $P=\{s|\ l(s)=pos\}$ and $N=\{s|\ l(s)=neg\}$. However, both of them contain wrong-labeled sentences which are affected by noise [5][44], so they propose to stick $P\backslash N$ together while keeping them separated. Hence, they proposed to apply stochastic gradient pairs (SGD) [26], and penalized relative distances for sentence triplets which greatly reduce undesirable moves.

$$L_{weak}=\sum_{<s1,s2,s3>} max(0,\ \lambda\text{-}dst(s_1,s_3)+dst(s_1,s_2)). \quad (3\text{-}1\text{-}1)$$

Where λ is the margin parameter, $dst(\cdot)$ is the Euclidean distance between sentences computed by their embedding layer representation:

$$dst(s_i,s_j)=||y_{si}-y_{sj}||_2. \quad (3\text{-}1\text{-}2)$$

and $< s_1, s_2, s_3 >$ denotes a valid triplet with $l(s_1) = l(s_2) \neq l(s_3)$. Eq.3-1-1 means the distance between same-label sentences s_1 and s_2 is required to be shorter than that between s_1 and s_3 with the opposite label by at least λ. Randomly choose P or N as the focus, then s_1 and s_2 are sampled from the target category, sentence s_3 is sampled from the other category. Figure 2 has shown the advantages of triplet-based training. The black nodes represent wrong-labeled sentences(noise), we can find that both training methods may generate undesirable moves, but triplet-based training method minimizes the influence of noise when sampled wrong-labeled sentences as figure 2(b) shows. What's more, it is useful that in triplet-based training, if the distances exceed the margin parameter λ, the derivative of L_{weak} becomes 0 which means no undesirable moves would happen at this situation.

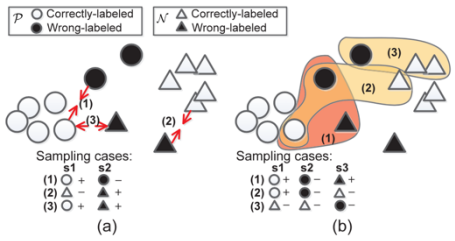

Fig. 2. Comparison between (a) pair-based training and (b) triplet-based training (Z. Wei et al. 2018)

Supervised Fine-Tuning. After obtaining a good enough sentence representation by the embedding layer, a classification layer is added on the top to further train the network using labeled sentences. The classification layer simply performs a standard affine transformation of the embedding layer output y, and then applies a *softmax* activation function to the result for label prediction.

Wei Zhao *et al.* compared WDE-LSTM and WDE-CNN on the ability of classifying, the results showed that WDE-LSTM performs slightly better than WDE-CNN when the labeled training set is sufficiently large. So, we choose WDE-LSTM as our baseline to enhance.

3.2 Model Enhancement – WDE-BiLSTM-Attention

The original WDE-LSTM network uses a unidirectional LSTM structure in the framework of weakly-supervised deep learning. The defect is that it cannot encode the information content of the current words according to the information below. For example: *"It is so boring!"*, *'so'* is used to describe *'boring'*, but when it comes to *"I am bored to death!"*, *'to death'* is used to describe *'bored'*. It may perform not so well in dealing with such kind of problems [29].

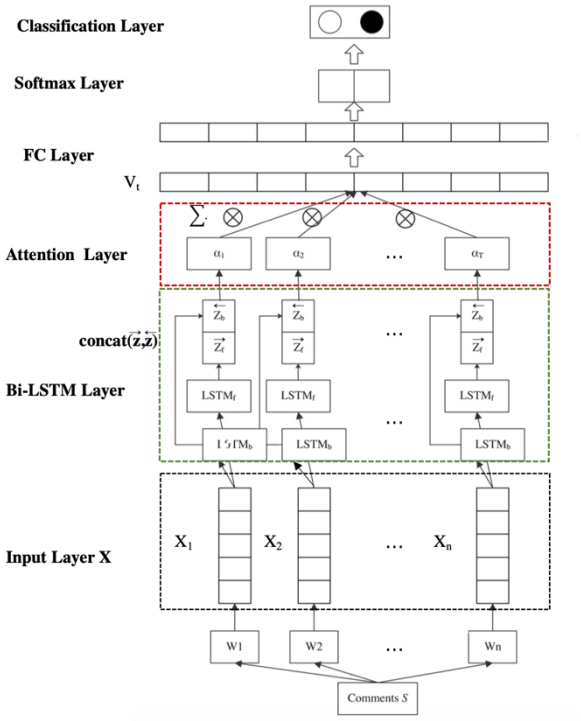

Fig. 3. Network architecture of WDE-BiLSTM-Attention

The improved WDE-LSTM model is mainly reflected in the network structure LSTM layer and attention mechanism. Figure 3 shows the network architecture of enhanced model. The attention mechanism is the abstraction of the phenomenon of human brains. In the early stage, it was widely used in the field of image processing. Currently, NLP also introduced this approach. According to the current word, different weights are assigned to the output matrix to generate a specific context representation. By calculating the probability distribution of attention, the key parts of things are given more weight to highlight, and then the model is optimized. Figure 4 is the text description of processing routine. The detailed explanations are shown following figure 4.

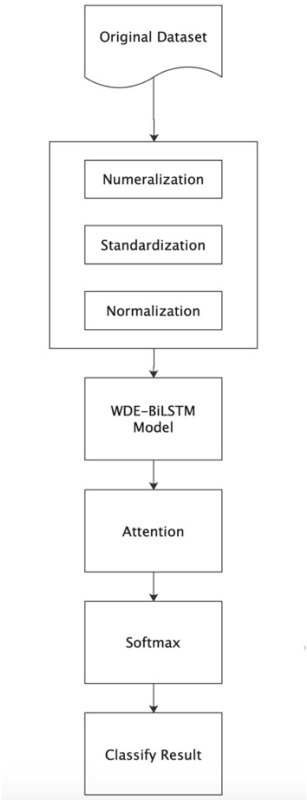

Fig. 4. Process routing planning

LSTM Layer. By building a bidirectional LSTM network to avoid semantic understanding errors, the LSTM layer contains the network of *forward* and *backward*, and then we connect the output. The output vector of words at time t in the statement after passing through the Bi-LSTM network is:

$$\overrightarrow{zt}=LSTM_f(x_t,\ \overrightarrow{z_{t-1}})$$
$$\overleftarrow{zt}=LSTM_b(x_t,\ \overleftarrow{z_{t+1}})$$
$$z_t=concat(\overrightarrow{z_t},\overleftarrow{z_t}) \quad (3\text{-}2\text{-}1)$$

Attention Layer. The attention layer is introduced to help capture important information in the text. It simulates that the human brain will assign more attention to the important content, while the unimportant part has less attention. Because the comment text is composed of multiple sentences, the importance of sentences in different positions will be different. Therefore, according to the characteristics of the attention mechanism:

$$u=sigmoid(w^TZ_t+b) \quad (3\text{-}2\text{-}2)$$
$$\alpha=softmax(u) \quad (3\text{-}2\text{-}3)$$
$$V_t=\sum^T{}_{t=1}\sum\alpha_tZ_t. \quad (3\text{-}2\text{-}4)$$

Among them, Z_t is the output of the Bi-LSTM layer, w is the vector that initialized randomly and continuously learned in training. b is the bias. $\alpha=[\alpha_1, \alpha_2,...\alpha_T]$ is obtained from u through the normalized exponential function *softmax*. V_t is the final output vector through the attention mechanism, which represents the feature information extracted from the comment text.

Softmax Layer. Through the attention mechanism, the output V_t passes through two FC layers, and finally transforms into two neurons. Each element value of the vector is limited to [0,1], which is normalized by the *softmax* function. Finally, a two-dimensional vector is output:

$$softmax(v_i)=\frac{e_i^2}{\sum_{j=1}^n e_j^2} \quad (3\text{-}2\text{-}5)$$

The input part of the model structure is the comment text. Through word segmentation and vectorization of the text, the fixed length feature vector of each word is obtained to extract the low-level features. The context information is introduced into the deep learning structure to realize the understanding from words to sentences. A weakly-supervised pre-training method is the core of the model. In order to obtain the emotional semantic embedding space of text sentences, the weakly labeled dataset is used to train the model. In the embedding space, the fixed length word vector is used to indicate that the words are closer in the embedding space, and the comments of different emotional categories are distanced from each other. On the basis of pre-training, the final classification model is obtained by training a part of a manually annotated dataset.

The final training process is divided into the following two steps:

Weakly-supervised pre-training. By using the weakly labeled dataset, the deep learning model based on LSTM is pre-trained with weak supervision, that is to say, the LSTM layer and the attention layer are fully trained to help capture the emotional semantic information of comment text statements and obtain the embedding space of semantic distribution.

Fine-tuning with supervised training. The model parameters trained in the first step are taken as the initial parameters, and then the full connection layer model is carried out by using the manually annotated dataset and classification layer connection behind the embedding layer. The supervised training of the whole model is carried out by using the manually annotated dataset, and the model parameters are fine-tuned to obtain the final emotion classification model.

4 EXPERIMENTS

We access 300,000 reviews under the top 15 wildlife parks from TripAdvisor. By eliminating repeated reviews, too short reviews, and special symbols, we finally got 274,773 effective reviews. The total number of words is more than 90 million. Then, we preprocess contents into word vectors to facilitate the subsequent import model for sentiment analysis. We randomly divide the preliminary processed comment data into 10,000 pieces as "manually annotated dataset" and the rest as "weakly labeled dataset". We mark positive comments as 1 and negative comments as 0 as table 2 shows.

Table 1. Rating System for TripAdvisor

Star Level	General Meaning
★	Terrible
★★	Poor
★★★	Average
★★★★	Very Good
★★★★★	Excellent

Table 2. Marking Categories Rules

Ratings	Category	Mark
1-3 stars	Negative	0
4-5 stars	Positive	1

Table 1 explains the meanings of one to five stars of TripAdvisor. We referred to this rating system. The tagging method in the weakly labeled dataset is to classify the sentiment according to the rating of the comment text by classifying those with more than three stars as positive comments and those less than or equal to three stars as negative comments, so as to obtain the weakly labeled dataset. A total of 195,753 positive comments and 79,020 negative comments were obtained.

The manually annotated dataset is to judge the emotional polarity of a text after reading a review by ourselves. For 10,000 randomly selected comment data, a total of 7,013 positive comments and 2,914 negative comments were obtained. The proportions were 70.13% and 29.14% respectively.

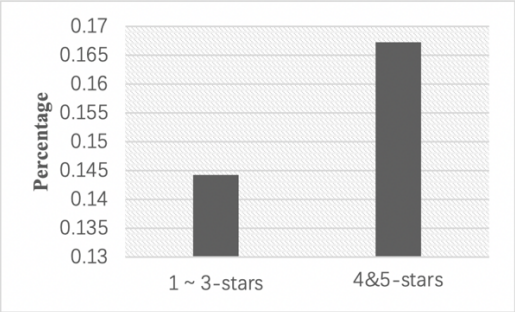

Fig. 5. Percentages of wrong-labeled sentences by ratings

Figure 5 illustrates that 14.73% of the 10,000 comments were classified incorrectly, including 16.73% of the positive comments and 14.42% of the negative comments.

The main reason is that the emotional category of the reviews cannot be corresponding to the star ratings. For example, a 5-star review generally expresses praise to the wildlife park, but contains negative words about other tourists, so it is automatically judged as a negative comment, but after reading it by ourselves, the comment is judged as positive. (Please refer to table 1 in appendix for more examples for this kind of comment noise.)

4.1 Oversampling

In this research, due to the imbalance between negative comments and positive comments, and usually in this type of research, negative comments are the focus of our attention. Therefore, negative comments are defined as positive cases in the confusion matrix, and the recall rate is more important than the accuracy rate. Figure 6 gives a visual representation. The red nodes are negative comments, they are far less than positive comments (the blue nodes) intuitively. We apply oversampling here to repeat negative reviews to better capture them. At the same time, the dataset is expanded with this method as table 3 shows.

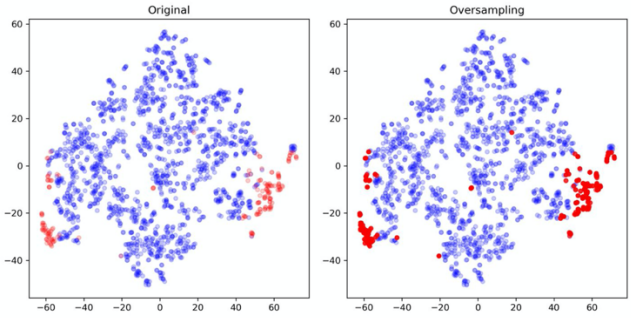

Fig. 6. Oversampling repeats negative reviews

In order to verify the performance of sentiment classifier of reviews, we usually use the following evaluation indexes: *Accuracy, Recall, F1 score and Precision*. If positive comments are chosen as positive cases in a confusion matrix, the recognition ability of negative comments will be seriously affected. When the recall rate in the sample is not balanced, the positive comment rate is taken as the negative evaluation index.

Table 3. Oversampling – Balance Two Categories of Reviews

	Emotion category	Number	Total
Weakly labeled dataset	Positive	195,753	391,506
	Negative	195,753	
Manually annotated dataset	Positive	7,013	14,026
	Negative	7,013	

4.2 Baselines and Comparison

The traditional classification algorithm is modeled. Sklearn machine learning library and Keras are used to train the commonly used classifier models: SVM, Naïve Bayes, RNN and LSTM. Then, the GridSearchCV is used to try various parameter combinations to determine the best ones. Based on TensorFlow framework, the **WDE-BiLSTM-Attention** model is implemented.

We firstly try different word vectorizing machines, in order to achieve the best performance of the subsequent sentiment classifier, we measure the advantages and disadvantages of Word2vec, CountVectorizer and

TfidfVectorizer by comparing the performance of the sentiment classifier.

Table 4. Performance Based on Different Word Vectorizer

Classifier	Precision	Recall	F1	Accuracy
(w2v) SVM	**0.79**	0.73	0.76	0.77
(w2v) NB	0.68	0.66	0.69	0.55
(w2v) LSTM	0.75	**0.83**	**0.83**	**0.79**
(cv) SVM	0.69	0.68	**0.69**	0.54
(cv) NB	0.61	0.60	0.54	0.51
(cv) LSTM	**0.70**	**0.69**	**0.69**	**0.66**
(tfv) SVM	0.63	**0.88**	**0.71**	0.55
(tfv) NB	0.60	0.62	0.66	0.62
(tfv) LSTM	**0.70**	0.72	0.69	**0.74**

From the table above, we can find that based on the same word vectorizer, LSTM has better performance in *recall rate* and *F1 score*. LSTM can extract the features of context in the memory and enhance the performance of the classifier, but SVM also has good performance in *precision*. Since the purpose of this paper is to identify negative comments more accurately, *the model with a higher recall rate and F1 score is the better model we need.*

Table 4 gives the results. Word2vec will not amplify the noise in the weakly labeled dataset, which will lead to the utilization of the LSTM characteristics and further enlarge the wrong information features, which will affect the final model classification performance.

Table 5. Classifier Performance Comparison

Classifier	Precision	Recall	F1	Accuracy	AUC
(w2v) SVM	0.7932	0.7341	0.7622	0.7746	0.79
(w2v) NB	0.6825	0.6684	0.6931	0.5534	0.67
(w2v) LSTM	0.7578	0.8384	0.8384	0.7925	0.88
(w2v)XGBoost	0.8184	0.8376	0.7982	0.8694	0.87
(w2v)CNN	0.7987	0.6897	0.8687	0.8844	0.87
(w2v)MemNet	0.8575	0.8372	0.8867	0.8465	0.85
(w2v)BiLSTM-Attention	0.8593	0.9044	0.8753	0.8768	0.86
(w2v)WDE-LSTM	0.8723	0.8878	0.8734	0.8726	0.86
(w2v)**WDE-BiLSTM-Attention**	**0.9157**	**0.9244**	**0.9112**	**0.9224**	**0.91**

The results in table 5 show that the **WDE-BiLSTM-Attention** model performs much better than other baseline classifiers on four indexes: *precision, recall, F1 score* and *accuracy*, which shows the model has great advantages in sentiment classification ability. Moreover, the influence of noise caused by a weakly labeled dataset on emotion classification effect can be overcome by manually annotated dataset fine tuning. Compared with the previous model, the enhanced model has a very obvious improvement in the results, which shows that the bidirectional LSTM can obtain the context information.

By introducing the attention mechanism, the words in different positions are given different weights, so as to achieve the function of identifying important information in the comment text, thus improving the performance of the model.

4.3 Sentiment Classification

(a)

(b)

Fig. 7. High frequency words in positive reviews (a) and in negative reviews (b)

We create word clouds of high frequency words in both positive and negative reviews, from which we can see that the keywords in positive reviews are: service, guidance, price, interpreter, tickets, (wild animal) species, etc., while the keywords of negative ones are: interpretation, welfare, protection, time guidance, time, fee, service, risk, etc.

We can notice from figure 7 that: both positive and negative reviews pay attention to service, interpret, guidance, ticket price, management, etc. Negative reviews show tourists' concern about the wildlife welfare and protection. The results preliminary showed that the tourists noticed the relationship between wildlife and human beings and realized the possible contradiction between environmental protection and wildlife tourism, as well as the safety problems caused by close human contact with wildlife.

Since we cannot explain the results more effectively only based on high frequency words, we will combine LDA and TF-IDF to extract the topic, in order to increase the interpretability, and further mine the topic.

4.4 Topic Mining Based on T-LDA

Due to the diversity and complexity of Internet information and the large amount of reviews, it will cost a lot of manpower and it is not efficient to understand tourists' attitude feedback on wildlife tourism by reading all the comments. Therefore, we propose to mine the main information in a large amount of review datasets through a topic model.

First of all, we segment the comments and filter important keywords through TF-IDF. Then, we train the LDA model with the obtained dataset and get the distribution between the topic and keywords under different sentiment classification. Through the analysis of the topic model, we can grasp the words under the corresponding topics, which helps us to interpret the tourists' attitude towards wildlife tourism to a certain extent.

Topic coherence evaluates the interpretability of the topic model and also the quality of the model. Therefore, we propose to take the measurement of topic coherence to determine the number of topics in the model. An appropriate K value can help us to provide meaningful and interpretable topics.

Positive Review Topic Analysis

According to the Coherence Score-K curve chart, the highest score is 0.650 when K=10 (Please refer to the figure 1 in appendix for more details). Therefore, we train the LDA model according to the selected K. Under each topic, we select the top 10 words according to the probability of the corresponding topic and regard the top words as the focus of the reviews. Because of the large number of topics, there will be repeated keywords under multiple topics, so we select the most meaningful 5 topics, which have good interpretability.

We can find from table 2 in appendix that the positive reviews of wildlife tourism can be summarized as follows: price, service, management, environment, transportation, time cost, wildlife variety and performance.

"I went there with my tour group, and indeed, it was incredible to see so many unique, interesting creatures of Mother Nature!"

"We spent a good amount of time at the center. We saw bisons, reindeers, wolves, and moose. We were in a group of 5 and we all had a blast! If you don't have much time to spend in Anchorage and want to see Alaska's animals, go here. A lot of the animals were rescued and habilitated. The gift shop had a lot of great Alaskan souvenirs. And the proceed goes towards the animals in the conservation."

"Great cause and educational. Worth a stop if you have time in your itinerary. Let's you get up close to the native animals that you'll only see from a distance in the wild."

The tourists who give positive comments think that the wildlife safari provides them with a chance to have close contact with wild animals that they cannot see normally. The wildlife here is safeguarded and well taken care of. It is an opportunity to provide science popularization education of nature. The time and price paid are appropriate and the cost performance is high.

In addition, we also found that most of the positive reviews list impressive wildlife names, which shows that tourists participating in wildlife tourism are very concerned about seeing as many species of animals as possible. Interpretation and service are also the concerns of tourists.

Negative Review Topic Analysis

According to figure 2 in appendix, the highest score is 0.878 when K=6.

We could find that in the negative reviews, the tourists' concerns are about the welfare of wild animals the most (Please refer to table 3 in appendix for more details). People were less concerned about price, environment and service than the positive reviews. Therefore, the negative attitude of tourists to wildlife tourism can initially reflect people's thinking about the ethical issues of animals and the environment. In addition, in topic 3, we can find the word "covid", which indicates that people's attitude towards wildlife has changed during the prevalence of COVID-19. People are aware of the negative impact wildlife may have on human beings.

"This attraction was not an attraction, it was a detraction. The animals were penned up, not in small zoo cages, but not in large spacious areas as we expected. In addition, it was clear that many of the animals were being bred to repopulate so that hunters could have the enjoyment of killing them. I would not recommend this attraction."

"It was so bad I don't even have words for it. We don't usually go to this kind of place but went under recommendation. I would like my money back because it surely was not used to help the animals."

"I know this place claims to protect and rehabilitate and it probably does to some degree, but what I witnessed were these beautiful bears pacing back and forth over and over again as if they were severely stressed in a cage. The poor fox was beautiful and I couldn't help feel sorry for him too. Wished I hadn't gone. I was depressed for the rest of my trip."

We found that in the negative evaluation, "regret", "pity", "cruelty", "sad" and other negative words with humanitarian tendency accounted for a large proportion. It can be seen that after participating in the whole tour experience, some tourists hold a negative attitude towards the wildlife park's treatment of animals.

To sum up, according to the analysis of topic mining results, the current wildlife tourism presents a very obvious polarization scene. The positive aspects of tourists mainly focus on the significance of science and educational popularization of wildlife tourism, which provides a sense of satisfaction for relatives and friends to travel together. The negative aspects mainly focus on the transportation, the lack of more extensive opportunities to observe wildlife, unreasonable ticket prices, improper management of the park, and the most important point: dissatisfaction with the park management's indifference to the environment and wildlife welfare. These aspects should be regarded as an important indicator of wider environmental problems and urgent need for proper management of tourism. Today, when COVID-19 is spreading around the world, if no positive measures are

token, it will continue to show negative downward slide of tourist satisfaction and animal welfare.

5 CONCLUSION

In this work, we proposed a novel deep learning framework named Weakly-supervised Deep Embedding BiLSTM-Attention Network based on classic WDE network, which change the original unidirectional transmission into bidirectional in the LSTM layer to capture the semantics in both directions, and the attention mechanism is introduced, which is helpful to capture the important information in the text and improve the accuracy of sentiment classification. The enhanced model WDE-BiLSTM-Attention has a notable improvement of classification accuracy and is efficient in the result, which shows that the bidirectional LSTM can obtain the context information. By introducing the attention mechanism, the words in different positions are given different weights, so as to achieve the function of identifying important information in the comment text, thus improving the performance of the model. It makes full use of a large-scale of weakly labeled information on the Internet and trains the model with a small-scale sample dataset, which greatly improves the accuracy and efficiency.

What's more, we focus on the ecological ethics contradiction of wildlife and human beings at this tourism turning point, from the perspective of tourists, extracting their opinions on wildlife. The negative aspects of tourists indicate the ethical thinking towards animals and environment. Today, when wildlife is so popular, people with humanitarian tendency is still fighting against this trend. Understanding the interaction between wildlife and tourists is quite forward-looking during this social background which COVID-19 is rampant.

5.1 Deficiency and Future Work

The results fully demonstrate the good performance of our emotion classifier in dealing with such problems, but there are still some deficiencies.

Due to the imbalance between positive and negative reviews in the original dataset, 4 and 5-star reviews account for a larger proportion. In order to balance, we set 3-star reviews as negative. However, 3 stars do not always mean a negative opinion, some reviews express positive tendency. We have to say that 3-star reviews are quite interesting objects worthy of study, and we can discuss the classification of 3-star reviews in the future.

Although oversampling improves the efficiency of parameter adjustment, it is possible to oversampling noise at the same time. If the error is multiplied, there is a risk of overfitting. In future work, we should consider how to imbalance positive and negative reviews more effectively while avoiding the influence of noise.

Emojis are removed in the process of segmentation. However, with the development of the Internet, more and more people use emojis to reflect their personal emotions. How to retain, effectively identify and extract the meaning of emojis are what we need to improve in the future.

In the process of topic model extraction, we found some meaningless words reflecting the theme, so we manually extracted and removed them to improve the effect of topic extraction. In future work, we can consider unifying different words with the same or similar semantics to mine important topic information with different expressions, so as to improve the effect of topic mining.

REFERENCE

1. R. Piryani, V. Gupta, and V.K. Singh. Movie Prism. A Novel System for Aspect Level Sentiment Profiling of Movies. Journal of Intelligent and Fuzzy Systems. (2017).
2. 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, 77: 236-246. (2017).
3. L. Chen, J. Martineau, D. Cheng, and A. Sheth. Clustering for Simultaneous Extraction of Aspects and Features from Reviews. NAACL-HLT, pages 789–799, (2016).
4. REN Jin-jin, LI Xue-yuan, SHENG Chen, and GAO Ying. Research on Attachment Behavior of Tourism E-commerce Websites Based on the Goal-Framing Theory. China Academic Journal Electronic Publishing House. (C), pages 107-108, (2012).
5. W. Zhao, Z. Guan, L. Chen *et al.* Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis [J]. IEEE Transactions on Knowledge and Data Engineering, 30(1):185-197, (2018).
6. Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. IEEE TPAMI, 35(8):1798–1828, (2013).
7. Y. Bengio. Learning Deep Architectures for AI. Foundations and Trends in Machine Learning, 2(1):1–127, (2009).
8. C.M. Bishop. Pattern Recognition and Machine Learning. Springer (2006).
9. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing from scratch. JMLR, 12:2493–2537, (2011).
10. A. Graves and J. Schmidhuber. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. Neural Networks, 18(5): 602–610, (2005).
11. D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. ACL, volume 1, pages 1555–1565, (2014).
12. T. Thet *et al.* Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards [Z]. London, England: SAGE Publications: 823-848, (2010).
13. M. Taboada, J. Brooke *et al.* Lexicon-Based Methods for Sentiment Analysis [J]. Computational Linguistics, 37(2): 267-307, (2011).
14. B. Pang, L. Lee, and S. Vaithyanathan. Sentiment Classification Using Machine Learning Techniques [C]. Proceedings of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press: 79-86, (2002).
15. D. Turney. Semantic Orientation Applied to Unsupervised Classification of Reviews [C]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 417-424, (2002).
16. D. Sanjiv and C. Mike. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. Proceedings of the 8th Asia Pacific Finance Association Annual Conference [J]. (2001).
17. Z. Fei, J. Liu, and G. Wu. Sentiment Classification Using Phrase Patterns [C]. The Fourth International Conference on Computer and Information Technology. CIT'04 IEEE. 1147-1152, (2004).

18. K. Dave, S. Lawrence, and M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews [C]. Proceedings of the 12th international conference on World Wide Web. ACM. 519-528, (2003).
19. C. Li, H. Wu, and Q. Jin. Emotion Classification of Chinese Microblog Text via Fusion of BoW and eVector Feature Representations [M]. Natural Language Processing and Chinese Computing Springer, Berlin, Heidelberg. 217-228, (2014).
20. W. Medhat, A. Hassan, and H. Korashy. Sentiment Analysis Algorithms and Applications: A Survey [J]. Ain Shams Engineering Journal. 5(4): 1093-1113, (2014).
21. B. Pang and L. Lee. Opinion Mining and Sentiment Analysis [J]. Foundations and Trends in Information Retrieval. 2(1-2): 1-135, (2008).
22. K. Dave, S. Lawrence, and D.M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW, pages 519–528, (2003).
23. S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6): 391, 1990. 22.
24. X. Ding, B. Liu, and P.S. Yu. A Holistic Lexicon-Based Approach to Opinion Mining. WSDM, pages 231–240, (2008).
25. L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu. Adaptive Recursive Neural Network for Target-Dependent Twitter Sentiment Classification. ACL, pages 49–54, (2014).
26. J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. JMLR, 12: 2121–2159, (2011).
27. R. Feldman. Techniques and Applications for Sentiment Analysis. Communications of the ACM, 56(4): 82–89, (2013).
28. X. Glorot, A. Bordes, and Y. Bengio. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. ICML, pages 513– 520, (2011).
29. A. Graves and J. Schmidhuber. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. Neural Networks, 18(5): 602–610, (2005).
30. H. Halpin, V. Robu, and H. Shepherd. The Complex Dynamics of Collaborative Tagging. WWW, pages 211–220, (2007).
31. S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8): 1735–1780, (1997).
32. M. Hu and B. Liu. Mining and Summarizing Customer Reviews. SIGKDD, pages 168–177, (2004).
33. N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A Convolutional Neural Network for Modelling Sentences. ACL, (2014).
34. Y. Kim. Convolutional Neural Networks for Sentence Classification. EMNLP, pages 1746–1751, (2014).
35. R. Kiros, Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-Thought Vectors. NIPS, pages 3294–3302, (2015).
36. D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate, arXiv: 1409.0473 (2014).
37. D. Tang, B. Qin, and T. Liu. Aspect Level Sentiment Classification with Deep Memory Network. Proceedings of the Conference on Empirical Methods on Natural Language Processing, pp. 214–224. (2016).
38. D.H. Pham and A.C. Le. Learning Multiple Layers of Knowledge Representation for Aspect Based Sentiment Analysis. Data Knowl. Eng. (2017).
39. J. Cheng, S. Zhao, J. Zhang, I. King, X. Zhang, and H. Wang. Aspect-Level Sentiment Classification with Heat (Hierarchical Attention) Network. Proceedings of the Conference on Information and Knowledge Management, ACM, pp. 97–106. (2017).
40. P. Chen, Z. Sun, L. Bing, and W. Yang. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. Proceedings of the EMNLP, pp. 452–461. (2017).
41. D. Ma, S. Li, X. Zhang, and H. Wang. Interactive Attention Networks for Aspect-Level Sentiment Classification. Proceedings of the IJCAI, pp. 4068–4074. (2017).
42. B. Liu. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, (2012).
43. B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. WWW, pages 342–351, (2005).
44. A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts. Learning Word Vectors for Sentiment Analysis. ACL, pages 142–150, (2011).
45. T.H. Nguyen and K. Shirai. PhraseRNN: Phrase Recursive Neural Network for Aspect-Based Sentiment Analysis. EMNLP, pages 2509–2514, (2015).
46. X. Qiu and X. Huang. Convolutional Neural Tensor Network Architecture for Community-Based Question Answering. IJCAI, pages 1305–1311, (2015).
47. A. Farman, K. Daehan, and K. Pervez. Transportation Sentiment Analysis Using Word Embedding and Ontology-Based Topic Modeling. Knowledge-Based Systems 174, pages 27-42, (2019).
48. R.Socher, B.Bengio, and C. Manning. Deep Learning for NLP. ACL(Tutorial Abstracts) :5, (2012).

APPENDIX

Table 1. Noise Caused by Non-Correspondence between Star Ratings and Reviews

Comments	Star ranking (weakly labeled)	Weakly labeled category	Manually annotated category
Paid $22 a person, all we got was an earful of sob stories and only got to see a few sick and injured animals. The tour guide talked and talked non-stop for over an hour. Most of the animals were hiding or sleeping and behind giant metal cages. Nothing about the tour was stimulating. If you want to donate, go ahead, but the tour itself felt like a complete waste of time.	5	1	0
Loved our visit to the centre and were shown around by lovely Becky. There were parrots, monkeys, snakes and sloths (my favourites)!! The staff all work really hard to keep the centre running. Would definitely go back and recommend to others.	2	0	1
Only place to see Baby sloths up close :) lots of other animals as well and they are really doing good by the animals. Tour is very informative!	3	0	1

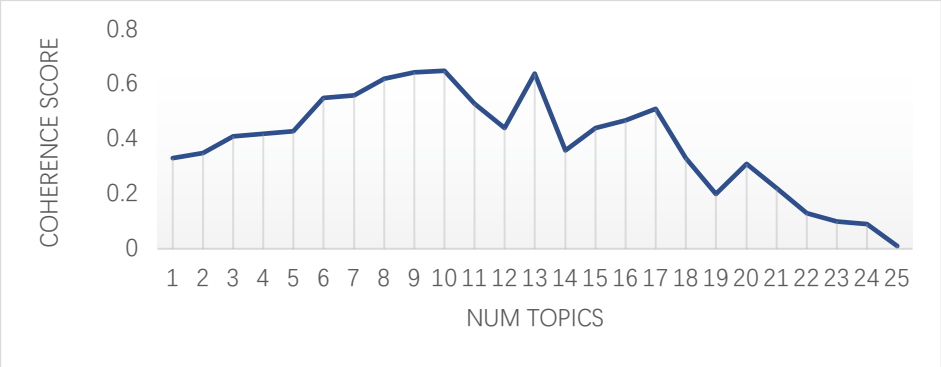

Fig. 1. Coherence-K value curve chart of positive reviews
It shows highest coherence score when K=10.

Table 2. Topic Words of Positive Reviews

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0	local*0.151	spend*0.082	enjoy*0.082	animals*0.192	service*0.067
1	guide*0.082	hrs*0.067	friends*0.082	wildlife*0.167	attitude*0.033
2	nature*0.132	see*0.200	want*0.045	experience*0.053	price*0.197
3	enjoy*0.082	drive*0.071	child*0.062	love*0.024	restaurant*0.045
4	visit*0.082	car*0.045	family*0.479	leisure*0.024	interpret*0.097
5	tour*0.082	take*0.081	worth*0.570	environment*0.12	staff*0.066
6	native*0.082	couple*0.374	together*0.047	rescue*0.037	nice*0.082
7	great*0.014	time*0.050	love*0.097	variety*0.071	food*0.076
8	kinds*0.014	bus*0.033	photo*0.048	clean*0.133	fee*0.053
9	recommend*0.083	fast*0.047	fun*0.098	protect*0.096	friendly*0.076

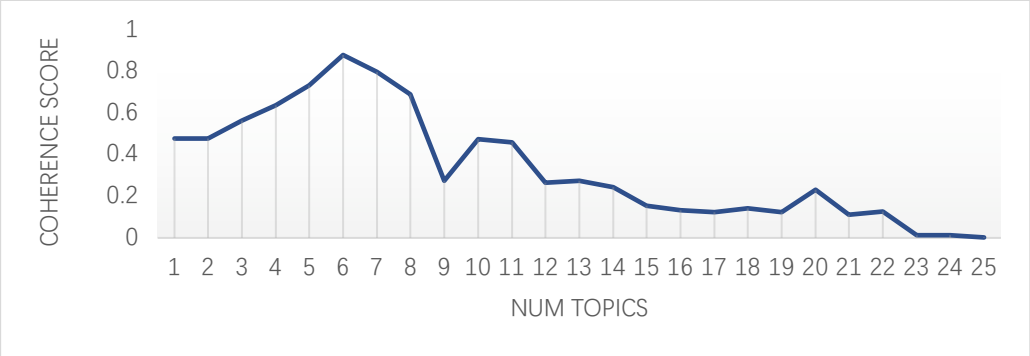

Fig. 2. Coherence-K value curve chart of negative reviews

Table 3. Topic Words of Negative Reviews

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
0	not*0.3683	time*0.2829	attitude*0.2906	animals*0.3499	price*0.3439
1	get*0.3656	short*0.2587	disappointed*0.2906	would*0.2982	high*0.3213
2	sad*0.2721	see*0.200	poor*0.2513	experience*0.2513	ticket*0.3195
3	never*0.2717	around*0.2446	child*0.2410	anti*0.2410	fee*0.2566
4	visit*0.2347	waste*0.2434	bear*0.2257	injured*0.2367	service*0.2001
5	regret*0.2311	take*0.2151	rude*0.2161	health*0.2350	money*0.1950
6	rude*0.2231	treat*0.1963	adverse*0.1961	close*0.2288	facilities*0.1938
7	horrible*0.2071	time*0.1963	same*0.1936	shelter*0.2170	giftshop*0.1916
8	old*0.1993	worthy*0.1909	care*0.1913	protective*0.1942	buy*0.1874
9	ever*0.1967	go*0.1862	covid*0.0023	suffer*0.1768	food*0.1785