Sentiment classification and visualization analysis of tourism comments: A canton tower example

Yanping Lin

College of Mathematics and Informatics, South China Agricultural University, Guangzhou, Guangdong, 510642, China

guaji_xd@stu.scau.edu.cn

Abstract. As calculated by the Ministry of Culture and Tourism, there were 308 million domestic tourist trips in China during the Spring Festival in 2023, witnessing a year-on-year increase of 23.1%. And the satisfaction of tourists with certain spots can be partly reflected in the comments and scores they made on social media. Therefore, this research was aimed at mining useful information from the comment and scores of Canton Tower. After collecting detailed comment information from the web, this research used the plot module of Python to make data visualization to observe the distribution of users' location, comment time, and comment label as well as the word cloud of remarks. Then the research used the data set to train three different sentiment analysis models including Naïve Beyas, SnowNLP, and Bert, then compared their accuracy in predicting. This research shows that over half of the comments came from Guangdong Province, most of the tourists were content with Canton Tower, and the number of comments has increased obviously since 2023. In addition, the research found that the model having the highest accuracy of sentiment analysis is the Bert model, about 90%.

Keywords: sentiment analysis, sentiment information classification, data visualization, online comment.

1. Introduction

As people's living quality is improved gradually and relevant policies have been announced and implemented, it shows a forward trend of the number of tourists in different spots in 2023, compared with recent years [1]. In order to acknowledge the tourists' gratification to those spots, the manager is paying increasing attention to the comment and scores on social media, which could have an influence on the tendency of other people's visit. Therefore, the sentiment analysis of comments could be an important issue [2,3].

The sentiment analysis of comments on the Internet has been a flourishing field of web information mining. In recent years, it has attracted wide attention from computer science, economics, and relevant disciplines [4,5]. Nowadays, most sentiment analyses are focused on movies' comments and commodities' comments [6]. However, there are few discussions about the remark of spots. Therefore, this research was aimed at analyzing and mining the data from Canton Tower's comments, in order to help managers to acknowledge the advantage and disadvantages of the spot, and therefore take measures to improve it.

The research mainly consists of two parts: data visualization and the comparison among sentiment analysis models. The project used a web crawler to get data on Canton Tower's comment information and make data visualization including comment time, comment label, and user's location through a Python library named Plotly. And three sentiment analysis models including Naïve Beyas, SnowNLP, and Bert were trained to analyze the data, and their predicting accuracy was then compared.

2. Method

2.1. Web data crawling

This research used selenium as the tool for data crawling. Selenium client initialed a service and started the Chromedriver through Webdriver. Then it sent an HTTP request to Chromedriver through remotewebdriver, the driver will receive and process the request and get the sessionid. After opening the port connected with the browser, the research took the browser as the remote server of webdriver. After opening the browser, each selenium script and each HTTP request were sent to the browser, the browser processed them and returned the result of it to a remote server. The remote server then returned the result to selenium scripts and wrote it into csv file called CantonTower.csv.

Some of the data gained from web has some features which are unfavorable in the following process of data visualization and model training. Therefore, this data will be under a successive preprocessing, including data clearing, word segmentation and the removement of stop words. The data gained from ctrip.com has some interrupting information like repeated and redundant comment and those with html labels and emoji. This kind of information was matched and cleaned through the regular expression, which can replace the html label with an empty string. The word segmentation is the basic part of data mining and information process, and "jieba word segmentation" is a word segmentation project of python community. This research collected common stop word files from the Internet and uploaded them into the word base. This research filtered those stop words from the comment data, which can exclude those interruption and improve the accuracy of word segmentation.

2.2. Data visualization

The research got detailed data of each comment from Ctrip.com, including user name, comment time, score, location, label as well as the content of comment. These information were written in a csv file called CantonTower.csv. Then the research employed the function of dataframe named value_count() to count the number of comment with different region and label. The counted statistics was used by a module of python called pyplot which is able to draw a pie chart to show the various proportion of different place and label. In addition, the number of comments in different date was also calculated and was used by the scatter model of plotpy_express library which displayed a scatter chart.

2.3. Naïve Bayes model

Assuming that there is any connection between each feature, and for each sample in the training data set, it includes dimensional features like $x = (x_1,\ x_2, \dots, x_n)$ as well as the class label set having different categories like $y = (y_1,\ y_2, \dots, y_n)$. For a new given sample x, the possibility of which category it belongs to can be calculated according to Bayes theorem [7]:

$$P(y_k|x) = \frac{P(x|y_k) \times P(y_k)}{\sum_k P(x|y_k) \times P(y_k)} \tag{1}$$

And the category with highest posterior probability is marked as predicted category like $argmaxP(y_k|x)$. Naïve Bayes makes an assumption on the independence of the conditional cumulative distribute. In other words, supposing that the feature of each dimension is independent, the conditional probability can be transformed like:

$$\mathrm{P}(x|y_k) = \mathrm{P}(x_1,\ x_2, \dots, x_n|y_k) = \prod_{i=1}^{n} P(x_i|y_k) \tag{2}$$

Then the formula of $P(y_k|x)$ can be transformed into:

$$P(y_k|x) = \frac{P(y_k)\times\prod_{i=1}^{n}P(x_i|y_k)}{\sum_k P(y_k)\times\prod_{i=1}^{n}P(x_i|y_k)} \tag{3}$$

Therefore, the classifier of naïve bayes can be expressed as:

$$f(x) = argmaxP(y_k|x) = argmax\frac{P(y_k)\times\prod_{i=1}^{n}P(x_i|y_k)}{\sum_k P(y_k)\times\prod_{i=1}^{n}P(x_i|y_k)} \tag{4}$$

Because all the value of the denominator of the formula are identical, so it can be ignored and the classifier of naïve bayes can be expressed as:

$$f(x) = argmaxP(y_k) \times \prod_{i=1}^{n}P(x_i|y_k) \tag{5}$$

This research gained data from the CantonTower.csv and signed for each sample, the mark given by users above or equal to 3 was marked as 1, and other mark below 3 was marked as 0. After labelling the data, this project separated the data set into training data set and test data set, added stop word and made word segmentation. Then the data was processed by TF-IDF to transform the text into vector. With the preprocessing finished, this project imported the Naïve Bayes model from the sklearn of python and choose the MultinominalNB as classifier, using data to train and test the model, then the accuracy was calculated and printed.

2.4. SnowNLP model

SnowNLP is a kind of Chinese natural language processing library which is based on Naïve Beyas algorithm [8,9]. After getting positive and negative samples, it will make word segmentation and exclude stop word from the sentence. After uploading Beyas model and using data to train it, the new model will be saved. The function called train is used to train to sentiment classifier and another function named classify is used to predict.

This project gained comment information data from the CantonTower.csv and signed for each sample, the mark given by users above or equal to 3 will be marked as 1, and other mark below 3 will be marked as 0. The preprocessed data was used to train the SnowNLP model to make sentiment scoring, then draw a nuclear density diagram to observe the distribution of score. Then the accuracy was calculated and printed.

2.5. Bert model

Essentially, Bert model is a two-stage model [10]. The first stage called pre-training, is similar to WordEmbedding, using language material without a mark to train a language model. And the pre-training includes two tasks, the Masked LM and the Next Sentence Prediction. The Masked LM will cover 15% words of a sentence at random, and predict what may the covered word means through its context. As many downstream tasks are based on the understanding of the relationship between two sentences, in order to enhance the understanding ability of the model, it will choose two sentences to determine whether the second sentence is the following text of the first sentence. And the second model is called fine-tuning, using preprocessed language model to complete relevant NLP downstream tasks. The Bert model will initial the parameter gained from the pre-training stage and use the labeled data from specific downstream tasks to fine-tune all parameters. Each downstream task has an independent fine-tuning model, though they use the same initialed pre-training parameter.

This project read comment information from the CantonTower.csv and marked each sample, the score given by users above or equal to 3 will be marked as 1, and another mark below 3 will be marked as 0. Given that the data set was big, this project disordered the set and reconstructed a new one. After reorganizing the data set, the comment sentence and label of it were taken out. With the imported BertTokenizer, the project added the encoded sentence to the list and covered it with an attention mask. Then the list was transformed into tensors. After designing the training, validation, and test dataset process, the project imported the text multi-classification model of Bert called Bert for sequence classification. Then it chose the optimizer to start the training process of data, after which the accuracy and loss were calculated and printed.

3. Result

3.1. Data visualization

3.1.1. The comment distribution of different regions. The pie chart, demonstrated in Figure 1, displays the distribution of the number of comments from different regions. It can be observed that over 50% of comment comes from Guangdong and the second is Shandong, about 9%.

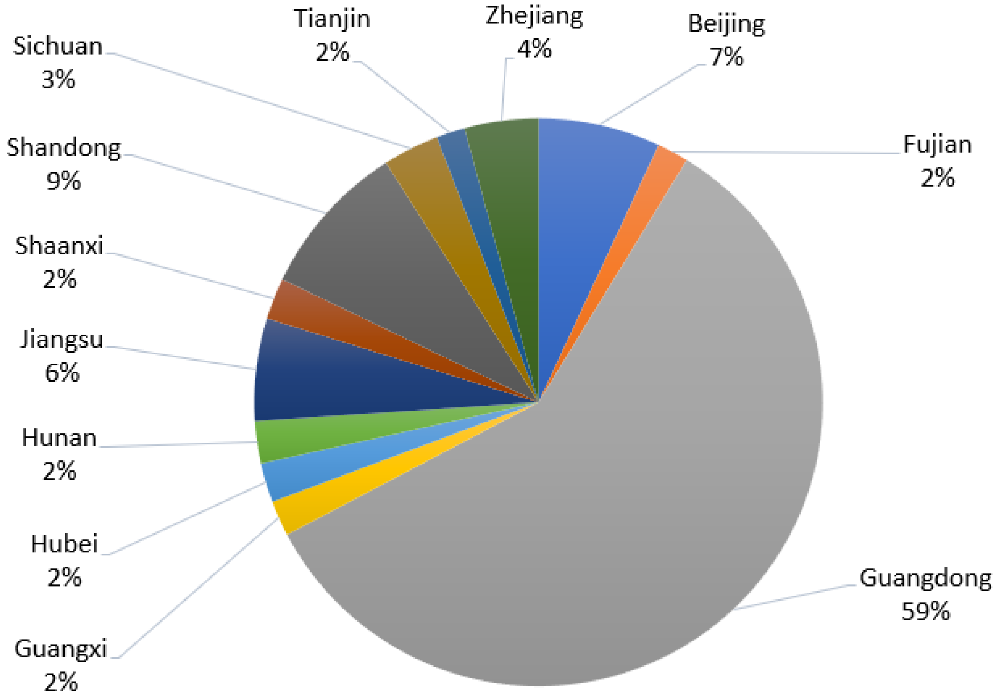

Figure 1. The distribution of comment locations.

3.1.2. The comment distribution of different labels. The pie chart, displayed in Figure 2, shows the distribution of different comment labels. It can be observed that most users are content with the CantonTower and about 53% of them thought it was perfect while 18% of them hold the belief that it was bad.

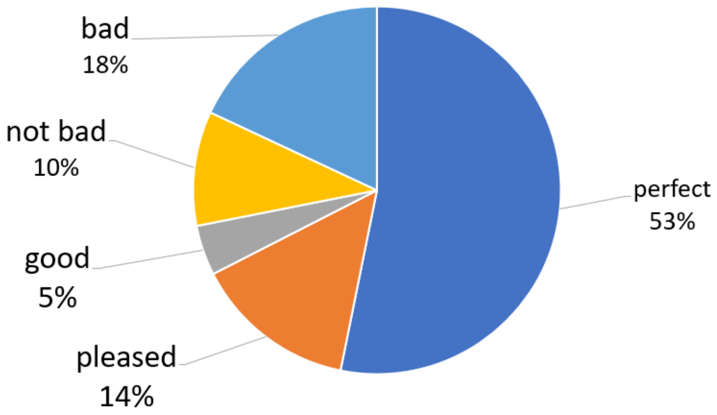

Figure 2. The distribution of comment labels.

3.1.3. The comment distribution of different time. The scatter chart, in Figure 3, conveys the information that most of comment was made in 2023, especially on January 27th, 2023.

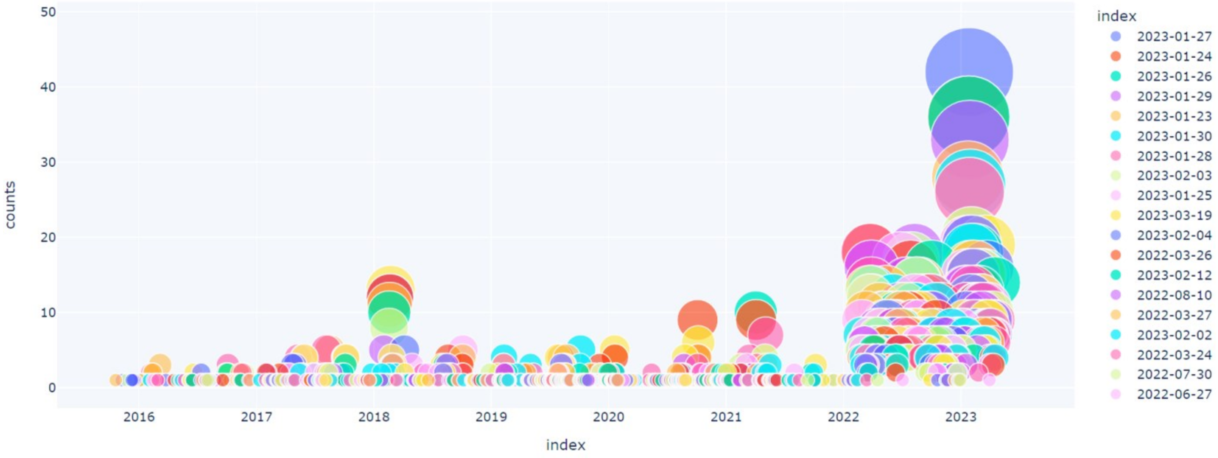

Figure 3. The distribution of various time.

3.2. Result comparison

3.2.1. Result of Naïve Bayes. The results are listed in Table 1. Through employing the function called score of *classifier* module, it would print the accuracy of the classification, which was about 79%. In the addition, utilizing the function called roc_auc_score, it could display the AUC value of classifier, which was around 79%.

Table 1. ACC and AUC performances of naïve bayes.

The accuracy of classification	79.24%
The AUC value of classifier	78.8%

3.2.2. Result of SnowNLP. Through employing the function called classification_report of dataframe module after training and testing. Results are demonstrated in Table 2. The function could print the detailed data of evaluating the model, like the precision of testing positive samples and negative samples, as well as the accuracy of model, which was about 84%.

Table 2. Classification results of SnowNLP.

	Precision	Recall	F1-score	support
0	0.78	0.70	0.74	1264
1	0.87	0.91	0.89	2751
accuracy			0.84	4015
Macro avg	0.82	0.80	0.81	4015
Weighted avg	0.84	0.84	0.84	4015

3.2.3. Bert. This research design two function called total_eval_accuracy and avg_val_loss to calculate accuracy of testing and the average loss of the model, which are shown in Table 3. After employing the total_eval_accuracy, the project printed the result was 90%, and the average loss during the testing process was 0.26.

Table 3. Results of bert model.

The accuracy of testing	90%
The average loss of testing	0.26

3.2.4. Result comparison. Drawing a conclusion from each model, this research formed a table as shown in Table 4. It can be observed from the table that the Bert model had the highest accuracy, which was about 90%. And that of the SnowNLP model was lower, about 84%, while the Naïve Beyas model was lowest, about 79%.

Table 4. Result comparison.

Model	Accuracy
Naïve Beyas	79%
SnowNLP	84%
Bert	90%

4. Discussion

Through training and testing each sentiment analysis model, the result shows that the Naïve Beyas model had the lowest accuracy, and the SnowNLP was 5% higher than it, while the Bert model had the highest accuracy of sentiment analysis. Their high accuracy of Bert could be explained by its training method. It masked certain words in the sentence and let the model predict those masked words, which could make a great achievement in semantic analysis. In addition, the model includes pre-training and fine-tuning, which can help the model be applied to different circumstances.

As this research mainly focused on the data of Canton Tower, there are still some limitations. In fact, the comment of different spots in China and even around the world could have great differences, which will affect the data visualization and the modelling process.

The result can be more general if more information on other spots could be collected, and the training process of different models could be improved. Therefore, the accuracy of predicting might be lifted.

5. Conclusion

This research found that there were an increasing number of visitors to Canton Tower in 2023, and almost half of them came from Guangdong Province while it also attracted tourists from other provinces. It can be observed from the pie chart that most of them were satisfied with the spot, about 53% of visitors commented on it with ‘perfect’. In addition, among three different sentiment analysis models, the Naïve Beyas model had the lowest accuracy, and the SnowNLP was 5% higher than it, while the Bert model had the highest accuracy of sentiment analysis, which could be attributed to the training method the Bert used. Bert model is a two-stage model, including the pre-training stage and fine-tuning stage.

This research provided more information about the sentiment analysis of comments about a spot. And it is recommended that the manager could use the Bert model to predict the comment of Canton Tower. Still, this research didn't take more information on other spots into account. In the future, it can be improved by collecting more comments from different spots around the world or the comment of the same spot from different websites or social media. Both the listed methods can promote the accuracy of the sentiment analysis model as well as the chart of data visualization.

References

[1] Seyfi, S., Hall, C. M., & Shabani, B. (2023). COVID-19 and international travel restrictions: the geopolitics of health and tourism. Tourism Geographies, 25(1), 357-373.

[2] Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment analysis of comment texts based on BiLSTM. Ieee Access, 7, 51522-51532.

[3] Alrumaih, A., Al-Sabbagh, A., Alsabah, R., Kharrufa, H., & Baldwin, J. (2020). Sentiment analysis of comments in social media. International Journal of Electrical & Computer Engineering 10(6), 2088-8708.

[4] Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment analysis of comment texts based on BiLSTM. Ieee Access, 7, 51522-51532.

[5] Yang, X., Xu, S., Wu, H., & Bie, R. (2019). Sentiment analysis of Weibo comment texts based on extended vocabulary and convolutional neural network. Procedia computer science, 147, 361-368.

[6] Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. International Journal of Computer Applications, 179(7), 45-49.

[7] Efron, B. (2013). Bayes' theorem in the 21st century. Science, 340(6137), 1177-1178.

[8] Chen, C., Chen, J., & Shi, C. (2018). Research on credit evaluation model of online store based on SnowNLP. In E3S Web of Conferences, 53, 03039.

[9] Lin, Y., Chen, L., & Zhang, C. (2022). Analysis of Tourist Hotel Impression Based on SnowNLP Model. In 2nd International Conference on Internet, Education and Information Technology, 373-378.

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.