Predicting Tourism Trend by Data Analysis

Mu Tian

13009535

1)AIMS

Nowadays machine learning is widely used in lot of fields, including tourism field, the most common case in tourism filed is using time series method to predict the future trend of tourism. However, as the influence of public opinion on society is increasing, the sentiment of people can also affect the tourism a lot. In this situation, sentiment analysis should be adopted in order to predict the impact of people's sentiments on data. Sentiment analysis is a machine learning method that can analysis and category the sentiment of sentences , so that the data can be used for further investigation. The aim of this project is to develop an approach to predict the tourism trend accurately using time series analysis as well as sentiment analysis of Facebook posts, so that the manager can have a better managing strategy on tourism. The overall aim can be broken into five parts, including 1)Collect related Facebook posts and the tourism trend data to set up a training set 2) Choose and deploy the classifier of sentiment analysis 3)Integrate data 4)Build the model 5)Validate and deploy the developed approaches.

2)BACKGROUND

With the increasing income of tourism industry, more and more people pay much attention to the trend of it since it can play a big role in the local industry’s income. So that there are more research on predicting the trend of tourism trend have been proposed to help the local companies to adjust their marketing strategy. The up-to-date method in predicting the trend of local tourism include using time series, which is also widely used in all kind of fields predicting the “future”, Yang Yang et al.(2019) gave insight into forecasting the tourism demand using Spatial-temporal, which predicted inbound tourism demand in 29 Chinese provincial regions. Also, other methods such as additive regression tree is also adopted, in one of the recent research, Chuanli Kang and Junfeng Gu (2019) used this machine learning method to analyse and build tourist flow forecasting model, in the article, the author mentioned that “Accurate prediction of tourist flow is a key issue in tourism economic analysis and development planning.”, and an additive regression tree model was build as a tourist flow prediction model. However, these analysis above have missed one important factor in affecting the tourism industry, that is the sentiment of people. For example, recently the frequent violent protesting activity in Hong Kong has led to a lot of negative comments on Hong Kong on social media, so that negative sentiments about Hong Kong have risen sharply among people. This sentiment resulted in a 'Significant fall-off in bookings' according to the report of BBC News(2019), which definitely has a great negative impact on the local tourism industry. According to the current situation, time series can’t accurately predict the trend of tourism all the time, and sentiment analysis can only judge how sentiments will affect the trend. So in this project we combined time series method and sentiment analysis to achieve a more accurate predicting trend, this trend can help the local tourism industry to adjust their managing as well as the marketing strategy, which can help them to ensure the tourism income.

3)RESEARCH PROJECT

□Significance of the project

Nowadays, with the improvement of peoples living standard, more and more people like to travel from time to time. This situation has made the role of tourism industry in local income increasingly important. Taking an example of Shanghai, China, the income of tourism industry of 2017 is 448.5 billion RMB, which accounts for more than 6.2% of the total GDP of Shanghai. Tourism industry nowadays is so important not only for local companies, but also to the local government. Local governments are promoting their cities in various ways to promote the development of their tourism industry. If the trend of local tourism can be accurately predicted, we can find out if the current marketing and managing strategy is good enough, helping us to locate the problem of current strategy, and to fix the problem, so that local tourism industry can be further developed . Also the prediction can help us to find when tourism is most likely to be at a low ebb in advance, then we can quickly put forward a coping plan. On the other hand, if the prediction of the tourism trend is not accurate enough, or there is even no prediction. The local tourism industry will not know what they can improve on the current strategy, and they can't properly deal with social sentiment since they have no idea about how sentiment will affect the tourism trend and its revenue. So that it is important to predict the tourism trend for it can help related industry and even local government to improve the marketing and managing strategy, thus rapidly improve the local tourism industry.

□Innovation of the project

The innovation of this project is adopting and combing multiple machine learning methods to predict the future trend of tourism. As mentioned above, the current up-to-data method is using time series to predict the trend. However, if there are some negative or positive reports about this city, social sentiment toward this city is very likely to be changed, which can have a huge impact on the trend of tourism. So what new about our project is that we use time series to predict the trend in normal situation, and then we also adopt sentiment analysis to include the impact of social sentiment in the prediction. In this way, we can have a more accurate prediction on the tourism trend. Also, if we are able to monitoring the mainstream social media continuously(eg. Facebook and Twitter), we will be able to have a real-time prediction on the trend of tourism. This will allow us to have an instant and appropriate reaction to the problem we may have in tourism, since we can compare it to the past index and the data is real-time updated.

□Briefly outline

In order to briefly outline the project so that it is easier to understand, the whole project can be broken down into several main tasks.

1.Collecting data and construct training set

In order to train the model accurately, we need to collect the relevant data. There should be two parts of data for the whole model. The first part should include the posts related to a certain city(Used as the keyword) on Facebook, and then date that it was posted. The second part should include the time(Most likely calculated on a monthly basis)and the tourism revenue(Used as the index of tourism trend) of that month. In order to focus on one certain city, the example city chosen here is Shanghai. The first part can be collected using web crawler tools, setting the keyword as the name of the city, which is in this case, Shanghai. The overall time period should be as long as possible to have a more convincing and reliable model. To collect the tourism revenue data, we can easily get it from relevant tourism websites(eg. 962020.com), the time period should also be as long as possible to make the time series more accurate and stable.

2.Choose and deploy the classifier of sentiment analysis

Compare classifiers of sentiment analysis and choose the most appropriate one. Use the classifier(eg. TextBlob) to calculate the score of the sentiment. Posts with scores above 0 are positive posts, scores below 0 are negative posts and scores exactly 0 are neutral posts. Record the score of each in the first data file.

3.Integrate Data

Calculate the mean value of the sentiment score monthly, and add a new column recording that value in the second data file. Then we have a complete training set including the time(Month), the monthly tourism revenue and the monthly mean value of sentiment score.

4.Build the model using Python

The model should be built similar to a time series model, however, the monthly mean value of sentiment score is also involved as a bias value of the model. Appropriate data analysis model can be built to find out how this bias value can affect the overall trend. In this way, the prediction can be made using the combination of time series and sentiment analysis.

5.Validate and deploy

When the model is built, it should validated before the actual deployment. The model should do the self-prediction and the future prediction test in real life. We can use the model to predict the tourism trend of next two months, and see if the predicted trend is accurate enough. If the prediction model is accurate enough, it can be deployed in real life. This project can be deployed in tourism industries. For example, travel agency can use this to determine and adjust their marketing strategy, and local government can use this project to help them managing the local tourism industry. However, if the accuracy is not good enough, the model should be further improved, get more training data or optimize the data analysis model, so that the reliability of the model can be guaranteed, and its value is further improved. EXTRA. Future improvement Since the time constraint is one year, there are works can't be done within time schedule, but worth being done in the future. That is the real-time prediction, in the future, if we can develop a real-time web crawler system, we will be able to monitor the social sentiment continuously. So that we can do the real-time tourism trend prediction, which can greatly help the relevant managing and marketing.

□Timelines

Time constraint of this project is one year, so all the work should be done within one year. The following part shows the rough estimation of time that each task will take.

1.Collecting data and construct training set

Based on the fact that this project requires large amount of data. For the first part of data, which is the post on Facebook, the collection of data should be done within 1.5 months . The second part of the data, which is the monthly tourism revenue, the collection of data should be done within 2 week.

2.Choose and deploy the classifier of sentiment analysis

This part needs to review possible methods of sentiment analysis and choose the best one. The procedure of listing and testing all the classifier should be done within 3 weeks. The deployment of the chosen classifier should be done within a week.

3.Integrate Data

This part needs to calculate the monthly mean sentiment value, and adding the value as a new column into the second part of the data(Tourism revenue part). The calculation of the mean value should be done within 5 days. The combining procedure should take no more than 3 days.

4.Build the model using Python

This part is the core part of the project, building a model that can predict the trend of tourism through time with the sentiment value as a bias value. This procedure contains collecting information of different models, choosing and combining different models. This procedure should be done within 3 months.

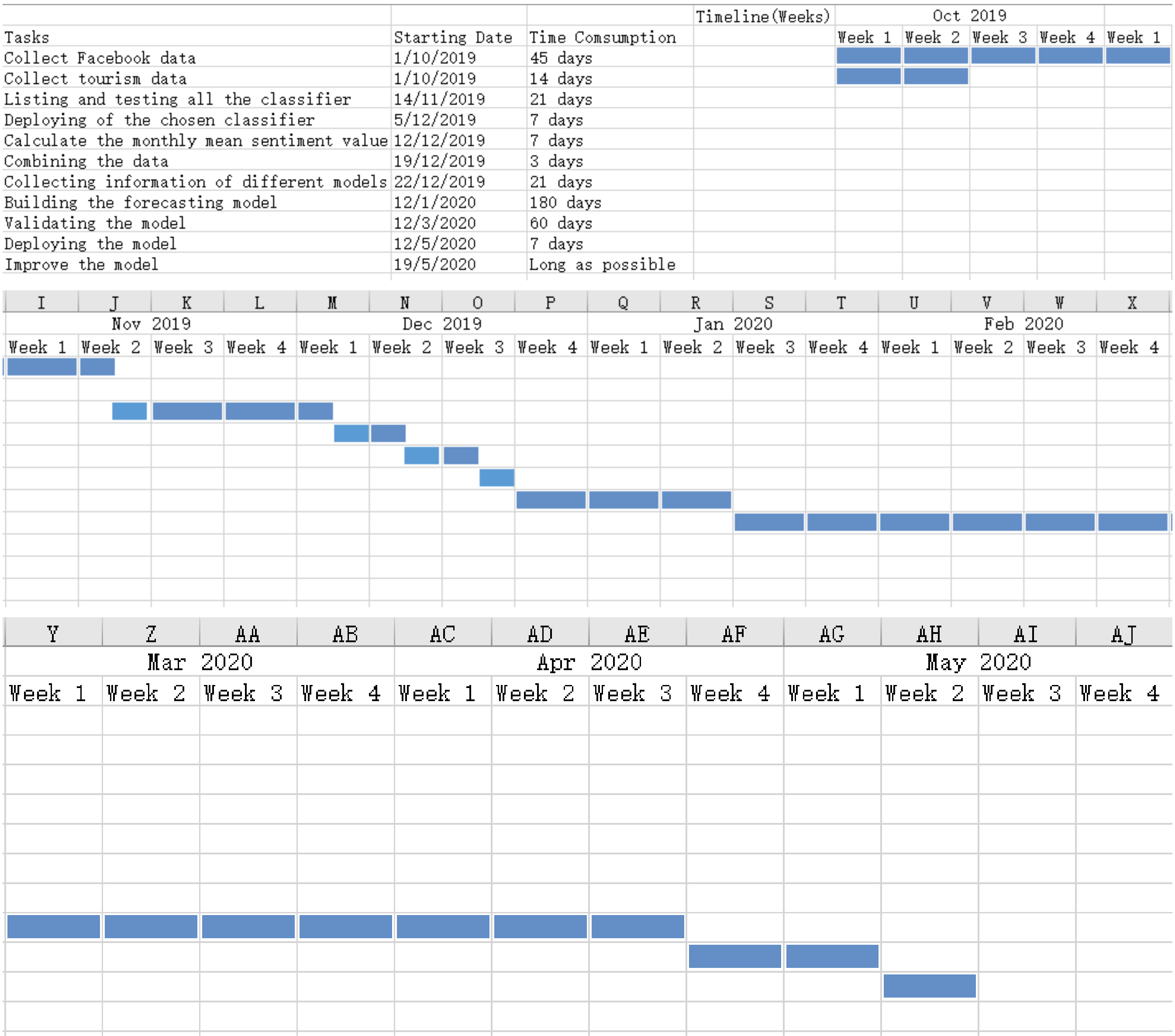
5.Validate and deploy

This part is to validate the accuracy of the model and finally deploy it into real life. Since the validate requires real life data for 2 months, the validate procedure should be no more that 60 days. The deployment procedure should take approximately a week.

6.Modifying

Since the time constraint is 1 year, there are still time left for this project. The time left should be used for further improvement on the model to acquire a better accuracy.

The following time in Gantt chart is based on the assumption that this project starts on 1st Oct 2019.



Tasks	Starting Date	Time Consumption
Collect Facebook data	1/10/2019	45 days
Collect tourism data	1/10/2019	14 days
Listing and testing all the classifier	14/11/2019	21 days
Deploying of the chosen classifier	5/12/2019	7 days
Calculate the monthly mean sentiment value	12/12/2019	7 days
Combining the data	19/12/2019	3 days
Collecting information of different models	22/12/2019	21 days
Building the forecasting model	12/1/2020	180 days
Validating the model	12/3/2020	60 days
Deploying the model	12/5/2020	7 days
Improve the model	19/5/2020	Long as possible

Timeline(Weeks)
Oct 2019
Week 1 Week 2 Week 3 Week 4 Week 1
I J K L M N O P Q R S T U V W X
Nov 2019
Dec 2019
Jan 2020
Feb 2020
Week 1 Week 2 Week 3 Week 4 Week 1 Week 2 Week 3 Week 4 Week 1 Week 2 Week 3 Week 4 Week 1 Week 2 Week 3 Week 4
Y Z AA AB AC AD AE AF AG AH AI AJ
Mar 2020
Apr 2020
May 2020
Week 1 Week 2 Week 3 Week 4 Week 1 Week 2 Week 3 Week 4 Week 1 Week 2 Week 3 Week 4

Fig1.Gantt Chart of the project

□Expected outcome

The expected outcome is that we can accurately predict the future trend of the tourism, so that we can have a better management over the local tourism. The model can adapt to data of different cities, so that is can be widely used. The investor can use this predicting model to help them supervise over the local tourism.

They can easily know whether their current managing and marketing strategy is good enough or need further improvement. When the current strategy is modified, the system can also help to predict the possible outcome of the modified strategy. Also, this system can help the investor know the current social sentiments toward this city, and how these sentiment will affect the tourism, so that the investor can react instantly according to the prediction. Overall, this project can help the investor to manage the tourism more easily and effectively, so that their tourism revenue can be improved steady.

4)BUDGET

The budget includes three main parts :software, hardware and human resource.

□Software

Software includes purchasing web crawler system, and other professional data analyzing tools. The approximate software budget will be 800$.

□Hardware

Hardware includes renting an office, purchasing PCs and relevant office tools. The approximate hardware budget will be: 30000 Office Renting + 800(PC) x 10 +2500(Office tools)=40500$

□Human resource

Human resource include the salary of the project team: 8500(per person per month) x 16 x 12 =1632000$

Total: So the total budget is approximately 800(Software)+40500(Hardware)+1632000(Human resource)= 1673300 So that the proposed budget should be around 1700000 $.

5)PERSONNE

The project team is consist of project manager, primary data analysts, senior data analysts, system testers and accountant. One project manager will be required to supervise and control the whole project, also he will be in charge of communicating with investors. Four system testers will be required to test the methods, including the sentiment analysis methods, model building methods and the overall forecasting system. Five primary data analysts will be required to collect useful data, three of them will be collecting the Facebook posts data, and two of them will be collecting the monthly tourism revenue data. Also the primary data analysts will help to build the model. Five senior data analysts will be required to build ,deploy and improve the data analysis model. One accountant will be required to record all the costs of the project.

REFERENCES

Yang,Y., Honglei,Z. 2018, 'Spatial-temporal forecasting of tourism demand', Annals of Tourism Research, vol. 75 March 2019, pp. 106-119.

Chuanli, K., Junfeng,Gu. 2019, ‘Analysis of Tourist Flow Forecasting Model Based on Multiple Additive Regression Tree’, IOP Conference Series: Materials Science and Engineering, viewed 27 September 2019, https://iopscience.iop.org/article/10.1088/1757-899X/490/4/042001/meta

BBC News 2019, Hong Kong protests: How badly has tourism been affected?, document, viewed 27 September 2019, https://www.bbc.com/news/world-asia-china-49276259