TEPM: traveller enrolment prediction mechanism using BERT-based feature clustering and LSTM models

Chung-You Tsai

Division of Urology,
Department of Surgery,
Far Eastern Memorial Hospital,
New Taipei City, 220216, Taiwan
and
Department of Electrical Engineering,
Yuan Ze University,
Taoyuan City, 320, Taiwan
Email: pgtsai@gmail.com

Ming-Yang Su

Department of Computer Science and Information Engineering,
Ming Chuan University,
Taoyuan City, 333, Taiwan
Email: minysu@mail.mcu.edu.tw

Christopher Chuang

Department of Computer Science and Information Engineering,
Tamkang University,
New Taipei City, 25137, Taiwan
Email: chrischuang@gmail.com

Chih-Yung Chang*

Department of Computer Science and Information Engineering,
Tamkang University,
New Taipei City, 25137, Taiwan
Email: cychang@mail.tku.edu.tw
*Corresponding author

Diptendu Sinha Roy

Department of Computer Science and Engineering,
National Institute of Technology,
Shillong, 793003, India
Email: diptendu.sr@nitm.ac.in

Abstract: The prediction of whether a tour group will form or not has a significant impact on travellers' future itinerary planning and travel agencies' control over hotel and flight bookings. Traditional methods rely solely on historical data, therefore lacks accuracy due to diverse tour attributes. The proposed mechanism, called TEPM, divides the enrolment prediction into three stages, including clustering, classification and prediction. Firstly, it clusters the tours to several groups according to the enrolment data. Secondly, natural language processing techniques are used to convert tour advertisements into feature documents. The BERT is employed to learn the relationship between advertisement feature documents and clusters. This enables the prediction of the group to which a given tour advertisement belongs. Finally, in the prediction stage, this paper employs dedicated LSTM models for each cluster to predict the number of enrolees. Experiments show that this approach performs well in terms of precision, recall, and F1 score.

Keywords: tour group prediction; feature clustering; natural language processing; BERT model; LSTM enrolment prediction.

Copyright © 2024 Inderscience Enterprises Ltd.

Reference to this paper should be made as follows: Tsai, C-Y., Su, M-Y., Chuang, C., Chang, C-Y. and Roy, D.S. (2024) 'TEPM: traveller enrolment prediction mechanism using BERT-based feature clustering and LSTM models', *Int. J. Ad Hoc and Ubiquitous Computing*, Vol. 46, No. 1, pp.14–26.

Biographical notes: Chung-You Tsai holds an MD from National Taiwan University and furthered his expertise with a PhD in Biomedical Informatics from National Yang-Ming University. In the medical specialty, he is a surgeon at Far-Eastern Memorial Hospital. At the same time, he is an Assistant Professor in the Department of Electrical Engineering at Yuan Ze University. He is dedicated to integrating technology with medicine. His specialisations include medical informatics and artificial intelligence, encompassing areas like AI, machine learning, and advanced biostatistics.

Ming-Yang Su received his PhD from National Taiwan University in 1997. He is currently a Professor in the Department of Computer Science and Information Engineering at Ming Chuan University, Taiwan. His research interests include deep learning, network security, intrusion detection/prevention, and wireless sensor networks.

Christopher Chuang received his BS in Computer Science from Pennsylvania State University in 2003 and Masters of Business Administration from MIT Sloan in 2008. He is currently a Board Secretary of Vanung University. He is also currently pursuing his PhD degree with the Department of Computer Science and Information Engineering, Tamkang University. His research interests include internet of things, blockchain, and AI.

Chih-Yung Chang received his PhD in Computer Science and Information Engineering from the National Central University, Taiwan in 1995. He is currently a Distinguish Professor with the Department of Computer Science and Information Engineering, Tamkang University, New Taipei, Taiwan. His current research interests include artificial intelligence, internet of things, wireless sensor networks, and smart healthcare. He has served as an Associate Guest Editor for several SCI-indexed journals.

Diptendu Sinha Roy received his PhD in Engineering from Birla Institute of Technology, Mesra, India in 2010. In 2016, he joined the Department of Computer Science and Engineering, National Institute of Technology (NIT) Meghalaya, India as an Associate Professor, where he has been serving as the Chair of the Department of Computer Science and Engineering since January 2017. His current research interests include software reliability, distributed and cloud computing, and IoT, specifically applications of artificial intelligence/machine learning for smart integrated systems. He is a member of the IEEE Computer Society.

1 Introduction

The prediction of traveller enrolment, particularly whether a tour group will form or not, holds an important role in travellers' itinerary planning and influences the precise control that travel agencies can exert over hotel and flight bookings (Sánchez-Medina and Eleazar, 2020). It is a dynamic and multifaceted challenge that significantly impacts the travel industry and the experiences of its customers. Given the distinctive characteristics and diverse attributes of tour groups, an integrated consideration of specific features such as travel duration, destination, holidays, layovers, attractions, pricing, and more becomes imperative (Cankurt and Subasi, 2015). Therefore, this paper focuses on the critical task of predicting the number of enrolees and the formation of tour groups, addressing the unique attributes and features of different tour groups.

Previous research in traveller enrolment prediction has predominately centred around two categories of predictive models: regression-based machine learning and deep learning prediction models (Silva et al., 2018). The regression-based machine learning models have offered valuable insights by examining historical enrolment data and establishing relationships between various features and enrolment outcomes. These machine-learning models have utilised traditional statistical techniques in forecasting (Kamel et al., 2008). However, machine learning models are limited in capturing complex traveller enrolment data, resulting in poor accuracy. On the other hand, deep learning prediction models, such as recurrent neural networks (RNNs) (Claveria et al., 2015b; Yao and Cao, 2020) and long short-term memory (LSTM) (Xia et al., 2022) networks, have emerged as a robust alternative. They excel in automated feature learning and accommodating the unique attributes and features of various tour groups. Nonetheless, deep learning models may require larger datasets to perform effectively.

While deep learning is generally better at feature extraction and establishing stronger relationships among different features for predictions compared to machine learning, most previous research has focused solely on modelling and predicting based on historical enrolment data. This approach may overlook the contextual information within travel advertisements. Typically, a traveller's decision to enrol in a particular tour is heavily influenced by the content of travel advertisements. These

advertisements contain various features related to travel arrangements, such as accommodations, restaurants, attractions, direct flights, departure and return times, which are significant factors affecting a traveller's decision to enrol in a tour.

Hence, this paper extracts features from travel advertisements and establish correlations and predictions based on whether travellers enthusiastically enrol in tours corresponding to these features. Moreover, different tour groups may have varying enrolment durations, which adds complexity to the clustering process. To address these challenges, the proposed TEPM extracts features from travel advertisements and group travel tours based on their enrolment data, ensuring that tours with similar enrolment trends are clustered together. This allows us to employ dedicated LSTM models to predict the number of enrolments for each cluster of tours.

Hence, this paper aims to propose a mechanism called traveller enrolment prediction mechanism (TEPM) using BERT-based feature clustering and LSTM Models to better predict travellers enrolment. The main contributions of the paper are summarised as follows:

- *Analysing advertisement features and predicting clustering:* Through in-depth analysis of advertising content, the proposed TEPM can identify the significant factors that influence whether tourists register. These factors are then used as features into the feature context construction and BERT (FCC-BERT) model to predict the clustering of travel groups. This helps segment potential customers into different groups, enabling more accurate predictions of tourist behaviour and decision-making.
- *Dynamic clustering based on registration deadlines:* The proposed TEPM uses the available enrolment data to dynamically cluster travel groups, reflecting variations in registration deadlines. This dynamic clustering method ensures that travel groups with similar registration schedules are grouped together, better adapting to changes in actual travel registration.
- *Data consistency alignment:* After completing the first-tier clustering, the proposed TEPM introduces crucial data alignment by standardising data. This ensures that all travel groups within the same cluster have the same number of registration days. This addresses the challenge of varying registration deadlines among different travel groups and contributes to more consistent data analysis.
- *Multilevel classification and prediction:* The proposed TEPM leverages FCC-BERT to extract these advertising features and associates them with predefined sub-clusters. Subsequently, it employs an LSTM prediction model to forecast future registration trends based on enrolment figures from the preceding days. This multilevel approach combines feature extraction, ad analysis, and prediction models, offering the potential for more intelligent travel group management and decision support.

The remainder of the paper is organised as follows. Section 2 compares previous related work. Section 3 describes the system model and problem formulation. Section 4 presents the detailed proposed TEPM. Section 5 provides performance study. The summary and future work are discussed in Section 6.

2 Related work

Traveller enrolment prediction has been an active area of research and received a great amount of attention in recent years. Machine learning and deep learning techniques have applied and offered valuable insights into this domain with varying degrees of success (Wang and Liu, 2022). Machine learning models, including support vector machines (SVMs) (Cankurt and Subasi, 2015), random forests, Gaussian Process Regression and gradient boosting have emerged as valuable tools for predicting traveller enrolment (Claveria et al., 2015a). For example, Cankurt and Subasi (2015) developed an SVM model using traveller demographic and search data that improved on logistic regression benchmarks. Similarly, Zhang et al. (2020b) found boosting algorithms outperformed SVMs and neural networks in predicting enrolments based on market campaign data. These models exhibit simplicity and computational efficiency, making them well-suited for applications with resource constraints. However, these studies have not considered advertising features and primarily focused on using SVM to analyse travellers' demographic information and market activity data. Additionally, they have not considered different enrolment durations for clustering. To comprehensively analyse travellers' enrolment behaviour, the proposed TEPM first groups travel tours based on different enrolment durations, which helps gain a better understanding of the impact of enrolment periods on traveller behaviour and the variations among different groups.

There are limitations of machine learning models, especially when facing complex data. The complex and dynamic nature of traveller behaviour presents significant challenges. These models may struggle with highly complex traveller enrolment data such as long-term temporal dependencies and delicate variations. They tend to fall short in effectively modelling long-term dependencies with their shallow neural networks (Law et al., 2019). Furthermore, their performance often depends on the quality of features that are manually crafted, and this process can be time-consuming and may not fully unlock the hidden potential within the data. These drawbacks emphasise the necessity for advanced and flexible methods, like deep learning models, to address some of the issues linked to machine learning in traveller prediction.

More recently various deep learning models using neural networks with multiple hidden layers have emerged as a powerful approach. Claveria et al. (2015b) developed a

deep network for enrolment prediction that extracted signals from raw search query data. Their model significantly outperformed shallow neural nets and other baselines. Bi et al. (2021) also found deep recurrent and convolutional neural networks are able to achieve high performance levels on a traveller enrolment benchmark task (He et al., 2021). Peng et al. (2021) used data quantified from social network with BERT model and built a tourism demand forecasting model based on gradient boosting regression tree as a forecast model. Deep learning models, prominently represented by LSTM networks, offer more advantages in capturing long-term dependencies with time series data (Abbasimehr et al., 2020; Zhang et al., 2020a). This feature is important in traveller prediction, where decisions and behaviours are influenced by both recent and distant historical experiences. LSTM networks excel in modelling and learning from these extended temporal patterns, enhancing the precision of enrolment forecasts. Also, deep learning models are well-equipped to analyse complex and nonlinear patterns within the data (Abbasimehr et al., 2020), which is more suited when dealing with the intricate and dynamic nature of traveller preferences and choices. LSTM networks outperform other models in revealing subtle connections among a multitude of features, thereby enhancing the precision of predictions (Law et al., 2019). However, deep models require large training sets and extensive hyperparameter tuning. Interpretability is also lower compared to some traditional techniques. However, they did not consider advertising features, the proposed TEPM employs a BERT model to analyse advertising features, aiming to comprehend the essential information contained in advertisements and explore how this information influences enrolment behaviour. This will assist us in gaining deeper insights into the potential influence of advertisements and provide more insights into travellers. By combining these diverse features and dimensions, the proposed TEPM is designed to offer more accurate and comprehensive analysis of traveller enrolment behaviour, potentially offering additional support and guidance for decision-making in the travel industry.

3 System model and problem formulation

This section introduces the definitions of notations, assumptions and constraints of this paper. Then, the problem formulation of this study is presented.

3.1 System model and notation definitions

This paper assumes that a travel agency has collected the dataset of past n tours, $n \geq 1$, denoted as $D^{past} = \{D_i^{past} | 1 \leq i \leq n\}$, where D_i^{past} is the data of the i^{th} tour information. Each tour data D_i^{past} is represented as $(A_i^{past}, [d_i^{start}, d_i^{end}], R_i^{past}, \pi_i^{past})$, where A_i^{past} denotes the tour advertisement, $[d_i^{start}, d_i^{end}]$ denotes the starting and ending dates respectively for sign-up, R_i^{past} denotes the enrolment data of tour D_i^{past}, and finally, π_i^{past} labels whether the i^{th} tour is successfully formed. Among these notations, the most complicated one is the enrolment data, R_i^{past}, which is further defined as follows. Let $g_{i,j}$ be the number of small group of people signed up on the j^{th} day of the i^{th} tour. Each small group could be family, friends, or a single individual. All of the small groups are denoted as the set $G_{i,j} = \{G_{i,j,k} | 1 \leq k \leq g_{i,j}\}$, where $G_{i,j,k}$ is the data of the k^{th} small group. Let r_i^{past} be the number of people enrolled on the j^{th} day of the i^{th} tour, it is obvious that the following expression (1) holds.

$$r_i^{past} = \sum_{k=1}^{g_{i,j}} |G_{i,j,k}| \tag{1}$$

This implies that we have $R_i^{past} = \{r_{i,j}^{past} \,|\, d_i^{start} \leq j \leq d_i^{end}\}$. For example, if $R_i^{past} = \{2, 4, 7, 1, 9\}$, which represents that the numbers of people enrolled by dates were 2, 4, 7, 1, and 9, respectively, assuming the days for registration were five for the i^{th} tour.

Given m tours which will be setup in the future. Let the m target tours be denoted as $D^{target} = \{D_i^{target} | 1 \leq i \leq m\}$, each tour data D_i^{target} has the same structure as $(A_i^{target}, [d_i^{start}, d_i^{end}], R_i^{target}, \pi_i^{target})$. Similarly, π_i^{target} labels whether or not the tour will be successful. Let $\mathcal{R}_i^{target} = \{r_{i,j}^{target} \,|\, d_i^{start} \leq j \leq d_i^{end}\}$ denote the set of $r_{i,j}^{target}$, where $r_{i,j}^{target}$ denotes the number of people signed up on the j^{th} day. If the number of enrolled people is not enough at the end date of registration, then the tour will be cancelled, otherwise it will be formed successfully. Let function F be a tour prediction function which is able to predict the registration information $\mathcal{R}_i^{target}$ for a given set D_i^{target}, $1 \leq i \leq m$. That is,

$$F : \left(D_i^{target}, \ldots, D_m^{target}\right) \rightarrow \left(\mathcal{R}_1^{target}, \ldots, \mathcal{R}_m^{target}\right). \tag{2}$$

3.2 Problem formulation and assumptions

This paper proposes a model to approach two goals. The first one is to predict every day's number of enrolments of a tour, i.e., r_i^{target} of R_i^{target}, $1 \leq i \leq m$ and $d_i^{start} \leq j \leq d_i^{end}$. The second one is to predict whether a tour can be formed or not, i.e., π_i^{target}, $1 \leq i \leq m$. This paper aims to find an optimal function $F^{opt} \in F$ which predicts the number of people signed up on the j^{th} day of i^{th} tour. The prediction error can be presented as the different of actual and predicted numbers of people signed up on the j^{th} day of i^{th} tour. That is, the prediction error is measured by expression (3).

$$r_{i,j}^{target} - \hat{r}_{i,j}^{target} \tag{3}$$

The overall error is measured by the error summation of each day during the registration-opened dates for m tours. Objective 1 reflects this goal. The objective function given in expression (4) describes a target function for predicting the number of registrations for future travel groups, m in

total. For each group, from the first day of registration opening d_i^{start} to the last day of registration d_i^{end}, the goal is to minimise the difference between the predicted daily registration numbers $\hat{r}_{i,j}^{target}$ and the actual registration numbers $r_{i,j}^{target}$. The mean squared error (MSE) is used here as the loss function to measure the discrepancy between the predictions and the actual numbers.

Objective 1: Find an optimal function $F^{opt} \in F$ to predict $\hat{r}_{i,j}^{target}$ of R_i^{target}, $1 \le i \le m$ and $d_i^{start} \le j \le d_i^{end}$, such that it satisfies:

$$\underset{F^{opt} \in F}{Min} \left(\sum_{i=1}^{m} \sum_{j=d_i^{start}}^{d_i^{end}} \left(r_{i,j}^{target} - \hat{r}_{i,j}^{target} \right)^2 \right) \tag{4}$$

The second goal of this paper is to predict whether the tour is success. Recall that π_i^{target} denotes whether the i^{th} tour is formed successfully. Let $\hat{\pi}_i^{target}$ denote the prediction of π_i^{target}. That is,

$$\hat{\pi}_i^{target} = \begin{cases} 1 & success \\ 0 & failure \end{cases} \tag{5}$$

Let $\varphi_i^{TP}, \varphi_i^{FP}, \varphi_i^{FN}$ denote the true positive, false positive and false negative, respectively, of the prediction that the i^{th} tour is successful ($\hat{\pi}_i^{target} = 1$). We have

$$\varphi_i^{TP} = \pi_i^{target} \times \hat{\pi}_i^{target} \tag{6}$$

$$\varphi_i^{FP} = \max\left(0, \hat{\pi}_i^{target} - \pi_i^{target}\right) \tag{7}$$

$$\varphi_i^{FN} = \max\left(0, \pi_i^{target} - \hat{\pi}_i^{target}\right) \tag{8}$$

The true positive φ_i^{TP} is held only when the prediction result is 'success' and it is really success. On the contrary, the false positive φ_i^{FP} is held only when the prediction is positive ($\hat{\pi}_i^{target} = 1$) but the actual answer is negative ($\hat{\pi}_i^{target} = 0$). The false negative φ_i^{FN} is held when the prediction is negative ($\hat{\pi}_i^{target} = 0$) but the label is positive ($\hat{\pi}_i^{target} = 1$).

For all tours, we have

$$\varphi^{TP} = \sum_{i=1}^{m} \varphi_i^{TP} \tag{9}$$

$$\varphi^{FP} = \sum_{i=1}^{m} \varphi_i^{FP} \tag{10}$$

$$\varphi^{FN} = \sum_{i=1}^{m} \varphi_i^{FN} \tag{11}$$

Let φ^{RC} and φ^{PC} denote the recall and precision of the m predictions. Therefore, the recall and precision is respectively presented as expressions (12) and (13).

$$\varphi^{RC} = \frac{\varphi^{TP}}{\varphi^{TP} + \varphi^{FN}} \tag{12}$$

$$\varphi^{PC} = \frac{\varphi^{TP}}{\varphi^{TP} + \varphi^{FP}} \tag{13}$$

Let notation φ^{Fscore} denote the F-measure. That is,

$$\varphi^{Fscore} = \frac{\left(\beta^2 + 1\right) * \varphi^{PC} * \varphi^{RC}}{\beta^2 * \varphi^{PC} + \varphi^{RC}} \tag{14}$$

The constant β in the *F-measure* indicates a ratio of recall to precision. In the case of $\beta = 1$, recall is as important as precision. That is,

$$\varphi^{Fscore} = \frac{2}{\frac{1}{\varphi^{PC}} + \frac{1}{\varphi^{RC}}} \tag{15}$$

In case that the precision is more important than recall, it implies that $0 \le \beta < 1$. Otherwise, we have $\beta > 1$. In the two extreme cases, $\varphi^{Fscore} = \varphi^{PC}$ as $\beta = 0$, and $\varphi^{Fscore} = \varphi^{RC}$ as $\beta = \infty$. Recall that notations π_i^{target} and $\hat{\pi}_i^{target}$ represent the actual and predicted outcomes, respectively, of whether the i^{th} travel group successfully launches. For such predictions, the F-score φ^{Fscore} is used to measure the effectiveness of recall and precision. Expression (16) is the second objective function in this paper, aiming to maximise the F-score for the prediction of whether the future m travel groups will successfully launch.

Objective 2: Find an optimal function $F^{opt} \in F$ to predict $\hat{\pi}_i^{target}$ $1 \le i \le m$, such that F^{opt} satisfies the following objective function.

$$F^{opt} = \arg\max_{F^{opt} \in F} \left(\varphi^{Fscore}\right) \tag{16}$$

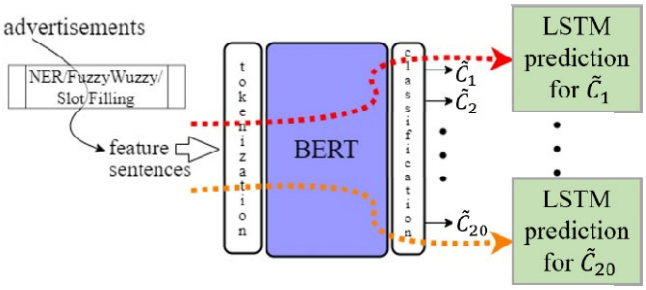
4 Proposed TEPM mechanism

The proposed model consists of three phases: Clustering phase, FCC-BERT phase, and LSTM prediction model construction (LSTM-PMC) phase. The first phase aims to cluster historical tours to several groups with different popularities according to enrolment data. The second phase is to train a BERT model which aims to predict the popularity of a tour from the advertisement context. In the third phase, a LSTM model is built to predict the number of enrolled people on the j^{th} day once the first $j - 1$ days' enrolment figures are known. The following presents the details of the three phases.

The entire proposed system is given in Figure 1, which consists of the three parts mentioned earlier in this section. In the training stage, we have n labelled tour samples as $(A_i^{past}, [d_i^{start}, d_i^{end}], R_i^{past}, \pi_i^{past})$, $1 \le i \le n$, where A_i^{past} is

the advertisement, $[d_i^{start}, d_i^{end}]$ is the starting and ending dates respectively for sign-up, R_i^{past} is the enrolment data, and finally, π_i^{past} indicates whether the i^{th} tour is successfully formed. The system proceeds as follows. The first part is a two-tier clustering algorithm to classify all advertisements A_i^{past}, $1 \le i \le n$, to m clusters, i.e., $\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_m$. The second part is an algorithm to extract features from advertisements, generate feature sentences, and train a classifier using BERT. Note that in the training stage to obtain a BERT-based classifier, the labels of tours are the cluster identifications, 1 to 20, obtained from the first part. The third part is an algorithm to build a LSTM prediction model for each cluster of the twenty ones.

Figure 1 Structure of the proposed system (see online version for colours)



After the three parts of the system have been completed in the training stage, it can be used to predict enrolment of a tour and thus know ahead if the tour will be formed successfully at the ending date of sign-up. As shown by the red (or orange) dashed line in Figure 1, the tour is classified to $\tilde{C}_1$ (or $\tilde{C}_{20}$) according to the BERT-based classifier with input feature sentences which are derived from its advertisement on webpages. Then the LSTM model trained for cluster $\tilde{C}_1$ (or cluster $\tilde{C}_{20}$) can be used to predict the enrolment of the tour once the first few days' enrolment data is gathered.

4.1 Two-tier clustering phase

The purpose of this phase is to partition the tour data into two-tier groups in which all tours in one group have similar popularities. The popularity is measured by the total of enrolled people from day to day. This phase includes two tasks: *data distribution task* and *two-tier clustering* task. The first task aims to normalise the tour registration data since the lengths of registration days of different tours are different. Then the second task aims to partition the tour data into two-tier clusters. The first-tier clusters are formed by partitioning the tour data according to the length of their registration days while the second-tier clusters are formed by further partitioning the tour in the same first-tier cluster into several groups according to their similarity in terms of popularity.

4.1.1 Data distribution task

This task consists of two steps. The first step is to cluster all tours according to the number of days for sign-up for the i^{th} tour. Then the second step will distribute the enrolment data such that all tours in the same clustering can be normalised.

Let the tours can be partitioned into φ clusters C_1, …, C_φ, according to their registration days. Let C_i^{days} be the number of registration days of cluster C_i. The enrolment data of a tour is counted by total number of signed up from day to day. Since tours have different registration days, the first step after manually clustering is to do alignment to make all the tours in the same cluster have the same registration days, i.e., all tours in cluster C_i have C_i^{days} figures in the records. Recall that d_i^{start} and d_i^{end} denote starting and ending dates, respectively, for sign-up of the i^{th} tour D_i^{past}. In the first step, the tour D_i^{past} will be grouped in cluster C_j^{days} if the following condition holds.

$$C_{j-1}^{days} \le \left(d_i^{end} - d_i^{start} + 1\right) \le C_j^{days} \tag{17}$$

After executing the first step, all tours can be clustered into φ clusters C_1, …, C_φ according to the number of days for sign-up.

Table 1 The algorithm of data distribution

Algorithm: data distribution mechanism	
Input:	$D_i^{past}, C_j^{days}, R_i^{past}$
Output:	Normalised enrolment data for each R_i^{past}
1	$idx = 0$;
2	$R_i^{past,extend}$ = null;
3	for each $r_{i,q}^{past}$ in D_i^{past};
4	$x_{i,q}^{past} = \frac{k \times C_j^{days}}{d_i^{end} - d_i^{start} + 1}$;
5	Repeat for $x_{i,q}^{past}$ times;
6	Append $r_{i,q}^{past}$ to $R_i^{past,extend}$
7	end for

The second step in the *data distribution task* aims to normalise the enrolment data for each tour. For each tour, the registration data for each day will be extended. The registration data will expand from the date range of the travel registration to the number of days corresponding to its cluster. Consider the registration data D_i^{past} of the i^{th} tour. Let the i^{th} tour D_i^{past} is in cluster C_j. Recall that notation $r_{i,q}^{past}$ denotes the number of people enrolled on the q^{th} day of the i^{th} tour. Assume that the registration data $r_{i,q}^{past}$ continues for k days in the enrolment data D_i^{past}. That is, there are no registrations for the subsequent k days, starting from the q-day of the i^{th} tour. Let $x_{i,q}^{past}$ a represent the

number of consecutive occurrences of value $r_{i,q}^{past}$ within a C_j^{days} date range. Then, the value of $x_{i,q}^{past}$ can be derived using the following formula.

$$x_{i,q}^{past} = \frac{k \times C_j^{days}}{d_i^{end} - d_i^{start} + 1} \quad (18)$$

Let $R_i^{past,extend}$ denote the normalised enrolment data extended from the original enrolment data R_i^{past}. Each $r_{i,q}^{past} \in R_i^{past}$ will be repeated $x_{i,q}^{past}$ times appeared in $R_i^{past,extend}$. The above-mentioned data normalisation will be applied for each $r_{i,q}^{past}$ in the enrolment data. Table 1 presents the data distribution mechanism.

Table 2 Original enrolment data R_i^{past} for a tour in cluster C_1

Tour_id	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7
XX001	0	0	18	18	18	41	41

Table 3 After normalisation obtaining $R_i^{past,extend}$ for a tour in cluster C_1

Tour_id	Day_1	Day_2	...	Day_{11}	Day_{12}	Day_{13}
XX001	0	0	0	0	18	18
Tour_id	...	Day_{28}	Day_{29}	Day_{30}	...	Day_{40}
XX001	18	18	41	41	41	41

The following gives an example to illustrate the normalisation task. In this study we have $m = 40$ and $r = 40$. Thus, all tours have been separated into four clusters, C_1, C_2, C_3, and C_4 orderly, while C_1^{days}, C_2^{days}, C_3^{days}, and C_4^{days} are 40, 80, 120, and 160, respectively. The purpose of data normalisation is to make every tour in C_i has $i \times 40$ days enrolled figures. For example, in Table 2 a tour in C_1 has only seven registration days, and the total signed up people is zero for the first two days, 18 for the next three days, and 41 for the last two days. After the normalisation operation, the tour's enrolment data of seven days has been extended to C_1^{days}, i.e., 40 days, as shown in Table 3. As shown in Table 2, it has $k = 3$ out of seven days keeping the same total enrolled 18 people. Therefore, the number of days keeping 18 enrolled people is expanded to $x_{i,3}^{past} = 17$ days among 40 days, where $x_{i,3}^{past}$ is derived by the following calculation.

$$x_{i,3}^{past} = \frac{k \times C_j^{days}}{d_i^{end} - d_i^{start} + 1} = \frac{3 \times 40}{7 - 1 + 1} = 17$$

4.1.2 Two-tier clustering task

After executing the *data distribution task*, all tours belonging to the same cluster C_j have the same length of registration days. The *two-tier clustering task* aims to further partition all tours in the same cluster into several smaller groups according to the similarity in terms of their enrolment data. Consider any pair of two enrolment data $R_x^{past,extend} \in C_j$ and $R_y^{past,extend} \in C_j$. Each $R_x^{past,extend} \in C_j$ is a multi-dimensional data which can be treated as a vector. Let $d(R_x^{past,extend}, R_y^{past,extend})$ denote the distance of two enrolment data $R_x^{past,extend}$ and $R_y^{past,extend}$. The distance can be obtained by applying expression (19).

$$d\left(R_x^{past,extend}, R_y^{past,extend}\right) = \frac{R_x^{past,extend} \cdot R_y^{past,extend}}{\left\|R_x^{past,extend}\right\| \times \left\|R_y^{past,extend}\right\|} \quad (19)$$

Table 4 The two-tier clustering algorithm

Algorithm: two-tier clustering	
Purpose:	Partition the tours into m clusters.
Input:	$(A_i^{past}, [d_i^{start}, d_1^{end}], R_i^{past}, \pi_i^{past})$,
Output:	m clusters, $\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_m$.
	// 1st-tier, clustering by registration days. //
1	Separate all tours to φ clusters C_1, C_2, …, C_φ in sequence according to their registration days, i.e., $d_i^{end} - d_i^{start}$.
	// 2nd-tier, clustering by cosine similarity and k-means. //
2	for each cluster C_i, $1 \le i \le \varphi$:
3	for every tour in the cluster:
4	Enrolment data alignment $R_i^{past} \rightarrow R_i^{past,extend}$
5	end for
6	Apply k-means to further partition tours in C_t according to cosine similarity
7	end for
8	Rename the m clusters to $\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_m$.

According to the distance defined above, clustering algorithms such as k-means or hierarchical clustering can be applied. Then all enrolment data in cluster C_j can be further partitioned into several smaller clusters, denoted as $C_{j,k}$. Let cluster C_j can be partitioned into n_j groups. There are totally

$$m = \sum_{j=1}^{n} n_j \quad (20)$$

Sub-clusters: The two-tier algorithm is given as shown in Table 4.

Upon the travelling tours have been partitioned into m clusters, all advertisements, A_i^{past}, $1 \le i \le n$, are used to train a BERT-based classifier.

4.2 The FCC-BERT classification phase

Given the training dataset, A_i^{past}, with labels which indicates each tour belonging to one of the m clusters. Let

$\tilde{C}_i$, $1 \le i \le m$, denote the m sub-clusters. The proposed method firstly extracts k features, i.e., feature set $F = \{f_1, \ldots, f_k\}$, for every tour from the advertisements. For example, the feature set includes date, price, destinations days and hotels, which might impact the decision of a traveller to join the group. Since the features implicitly exist in the advertisement, the techniques named entity recognition (NER) and fuzzy wuzzy are further applied to extract features from the advertisements which were appeared in webpages. Moreover, the slot filling is also applied to identify whether or not each feature exists in the advertisement, as shown in Figure 2(a). An instance of feature sentences of a tour is likely to be 'departure date is 6 April 2022. Destinations (countries or cities) are D_1, D_2, and D_3. Attractions are A_1, A_2, and A_3. The travel expense is about USD2,000. Flight company is F_1 and takes off at 5 PM. Travel agency is T. Days of the journey is 12. Hotels are H_1, H_2, H_3, H_4, and H_5. Optional tours include O_1 and O_2.

Figure 2 The proposed method to get a classifier from advertisements, (a) from advertisements to feature sentences (b) using BERT to train a classifier (see online version for colours)

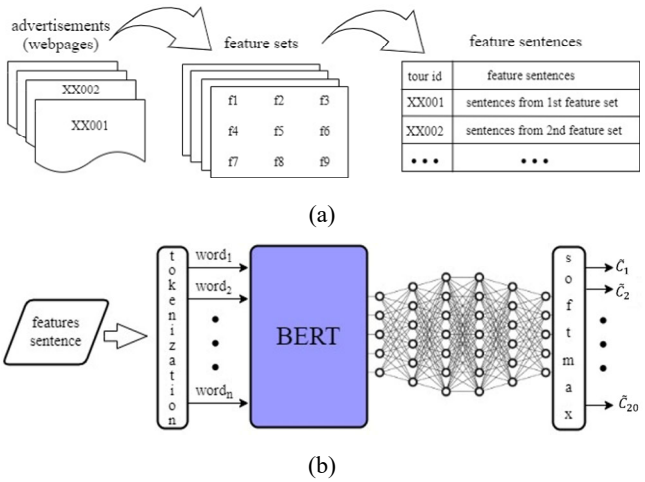


Figure 2(a) depicts the pre-processing task which extract the predefined feature sets from the advertisement. After feature extraction from the advertisement, all the feature sentences are taken as inputs of the bidirectional encoder representations from transformers (BERT), which is a well-known pre-trained language model. The downstream task of the BERT is a classifier which aims to extract the features of the advertisement and output one class of $\tilde{C}_i$ which is the most related to the advertisement features. The architecture design of the travelling group classifier is given in Figure 2(b).

The BERT is a state-of-the-art natural language processing model developed by Google. It utilises a bidirectional transformer architecture to pre-train on the vast amounts of text data, enabling it to understand context and semantics in a bidirectional manner. BERT has achieved remarkable performance on a wide range of NLP tasks, making it one of the most influential language models in the field. One of BERT's key innovations is its bidirectional processing. Unlike traditional NLP models, BERT can simultaneously consider both the left and right context for a word, allowing it to better understand the meaning of words in different contexts and thus improving its performance on a wide range of NLP tasks.

Let S denote the sentence that combines the k features. The first step is the tokenisation which partitions S into ω tokens, say $s_1, \ldots, s_\omega$. Then the ω tokens are converted into embedding vectors using a pre-trained word embedding matrix $E = [e_1, \ldots, e_\omega]$, where e_i is the embedding vector for token s_i. Let $P = [p_1, \ldots, p_\omega]$ be the positional embedding, where p_i is the positional embedding vector for position i. Additionally, the matrix P is added to the embedding vectors to capture the relative positions of the ω tokens. Let $X = [x_1, \ldots, x_\omega]$ denote the combination of embedding vectors E and positional embedding P. The BERT contains multiple transformer encoder layers, each with self-attention mechanisms and feed-forward neural networks. The input sequence X will go through these encoder layers, resulting in layer-wise output H_i of the i^{th} layer. As its input and then perform the self-attention operations. For each encoder layer i, the Q, K and V matrices are calculated as follows.

$$Q_i = XW_i^Q,\ K_i = XW_i^k \text{ and } V_i = XW_i^Q \tag{21}$$

where W_i^Q, W_i^k and W_i^V are learnable weight matrices for the query, key, and value projections in the i^{th} encoder layer. Then the multi-head attention mechanism is applied to these Q, K, and V matrices for each encoder layer:

$$\begin{aligned} H_i &= multiHead\left(Q_i, K_i, V_i\right) \\ &= Concate\left(head_1, \ldots, head_h\right)W_i^O \end{aligned} \tag{22}$$

where $head_j$ represents the j^{th} head of the multi-head attention mechanism, and W_i^O is the output weight matrix for the i^{th} encoder layer.

BERT uses the representation of the [CLS] token from the final encoder layer as the aggregate representation of the entire input sequence. Mathematically, this can be expressed as

$$C = H_L[\theta], \tag{23}$$

where L is the index of the final encoder layer, and θ denotes the [CLS] token's representation. The representation C is fed into a fully connected layer. The fully connected layer includes weight matrix W and bias matrix b. Finally, the model's prediction is obtained as

$$\hat{Y} = softmax(CW + b), \tag{24}$$

where $\hat{Y} = [\hat{y}_1, \ldots, \hat{y}_m]$, and each $\hat{y}_m$ represents the probability of belonging to class $\tilde{C}_i$.

The downstream task typically uses the cross-entropy loss to compute the difference between the predicted results and the actual labels. The actual labels are represented as $Y = [y_1, \ldots, y_m]$, where each y_m is either 0 or 1, indicating if the label belongs to class $\tilde{C}_i$ or not. The loss function is

$$L = -\sum_{i=1}^{m} y_i \log(\hat{y}_i), \tag{25}$$

which is the cross-entropy loss, used to access the disparity between the predicted classification results and the true labels.

The algorithm of FCC-BERT is shown in Table 5.

Table 5 The algorithm of FCC-BERT

Algorithm: FCC-BERT	
Purpose:	Train a BERT-based classifier.
Input:	A_i^{past}, $1 \le i \le n$, and their cluster identifications as labels.
Output:	A BERT-based classifier.
	// Derive features from every advertisement and use them to generate feature sentences. //
1	For each advertisement:
2	Apply NER to get tokens with attributes, and then use fuzzy wuzzy to remove similar ones.
3	Apply slot filling to get features and generate feature sentences from them.
4	end for
5	Construct a deep learning model using BERT as word embedding and DNN plus FNN as a downstream task for a classifier.
6	Use feature sentences instead of advertisements as input and cluster identifications as labels to train the model.

4.3 The LSTM-PMC phase

Once the BERT classifier is obtained, it can be used to classify a tour to one of the total m clusters for an input of the advertisement. However, we still need a LSTM prediction model for each cluster to predict the enrolment if a few days of numbers of people enrolled are given. That is, given the first ρ days (Day_1, Day_2, …, Day_ρ) enrolled numbers, the LSTM model aims to predict the next day's ($Day_{\rho+1}$) enrolment. Continuedly, with the ρ days (Day_2, Day_3, …, $Day_{\rho+1}$) enrolled numbers, the LSTM model can predict the enrolment of $Day_{\rho+2}$. Therefore, the travel agency has a chance to know in advance if a tour can be formed successfully at day k just according to the enrolments of the first ρ days, including $Day_{k-\rho+1}$, Day_3, …, Day_{k-1}.

For instance, in Figure 3, $\rho = 6$ and according to the advertisement of the tour, it is classified to $\tilde{C}_3$. Once the LSTM for $\tilde{C}_3$ is built, we can apply the model to predict enrolment of day 7, given the previous six days' enrolment data. Similarly, we can predict enrolment of day 8, given the previous 6 days' (day 2~day 7) enrolment data. In other words, at day 6, we can supposedly predict if the tour group will be formed successfully at the day when the registration is closed.

To conduct the training process, the travelling data should be pre-processed first. Recall that

$$R_i^{past,extend} = \left\{ r_{i,j} \middle| 1 \le j \le C_k^{days} \right\}, \tag{26}$$

denotes the enrolment data of the i^{th} tour D_i^{past} where D_i satisfies condition

$$C_{k-1}^{days} \le \left(d_i^{end} - d_i^{start} + 1 \right) \le C_k^{days}. \tag{27}$$

The training dataset X_{train} should be prepared by

$$X_{train} = \begin{bmatrix} r_{i,1} & \cdots & r_{i,\rho} \\ r_{i,2} & \cdots & r_{i,\rho+1} \\ \cdots & \cdots & \cdots \\ r_{i,C_k^{days}-\rho} & \cdots & r_{i,C_k^{days}-1} \end{bmatrix} \tag{28}$$

While the label Y_{train} of X_{train} should be prepared by expresssion (29).

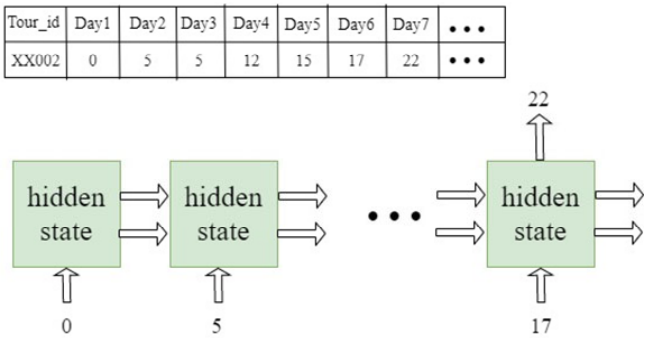
$$Y_{train} = \begin{bmatrix} r_{i,\rho+1} \\ r_{i,\rho+2} \\ \cdots \\ r_{i,C_k^{days}} \end{bmatrix}. \tag{29}$$

The LSTM takes a sequence of travel enrolment data for the previous ρ days, denoted as $r_{i,1}$, …, $r_{i,\rho}$, as inputs and aims to predict the number of travel enrolment for the next day, denoted as $r_{i,\rho+1}$. Initially, the input gate decides the value i_t based on the current input $r_{i,k}$ and the previous hidden state $h_{i,k-1}$. It is calculated using the following formula:

$$i_t = \sigma\left(W_i \cdot \left[r_{i,k}, h_{i,k-1} \right] + b_i \right) \tag{30}$$

where W_i is the weight matrix, σ is the sigmoid activation function, $r_{i,k}$ is the current input, $h_{i,k-1}$ is the previous hidden state, and b_i is the bias.

Figure 3 A LSTM prediction model for $\tilde{C}_3$ (see online version for colours)



The purpose of input gate is to determine what new information should be added to the cell state. Then the forget gate determines what information from the long-term memory should be kept. It is calculated using the following formula:

$$f_t = \sigma\left(W_f \cdot \left[r_{i,k}, h_{i,k-1} \right] + b_f \right) \tag{31}$$

where W_f is the weight matrix and b_f is the bias. The forget gate helps the LSTM to remember or forget information

from the past. On the other hand, the cell state C_t is updated by combining the previous cell state with the new candidate values. This is done using the following formula:

$$C_t = f_t * C_{i,k-1} + i_t * \tanh\left(W_c \cdot \left[r_{i,k}, h_{i,-1}\right] + b_c\right) \tag{32}$$

where W_c denote the weight matrix, and tanh is the hyperbolic tangent activation function. This step controls the information that is stored in the cell state.

The output gate determines the hidden state $h_{i,k-1}$ based on the updated cell state. It is calculated using the following formula:

$$o_t = \sigma\left(W_o \cdot \left[r_{i,k}, h_{i,k-1}\right] + b_o\right) \tag{33}$$

where W_o is the weight matrix and b_o is the bias. The output gate controls the information that is passed to the final prediction. The algorithm of LSTM-PMC is shown in Table 6. Upon the m models $LSTM_i$, $1 \le i \le m$, have been trained completed based on the pair of (X_{train}, Y_{train}), the m models can be applied to predict the enrolment data.

Table 6 The algorithm of LSTM-PMC

Algorithm: LSTM-PMC	
Purpose:	Build a LSTM model for each cluster
Input:	Enrolment data of m clusters, $\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_m$.
Output:	m LSTM models.
1	for each cluster $\tilde{C}_i$;
2	Let $R_i^{past,extend} = \{r_{i,j} \mid 1 \le j \le C_k^{days}\}$
3	Prepare X_{train} training data according to expression (28)
4	Prepare Y_{train} labels according to expression (29)
5	Construct a $LSTM_i$ model with ρ time sequence.
6	Train the $LSTM_i$ model based on (X_{train}, Y_{train})
7	end for

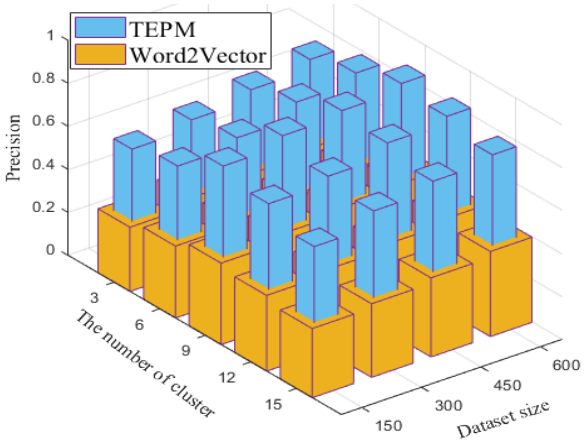
5 Performance study

This section evaluates the performance enhancement of the proposed TEPM in comparison to linear and LSTM models. The proposed TEPM leverages BERT and LSTM for predicting tour enrolment. The subsequent sections provide details regarding the simulation environment and present the results.

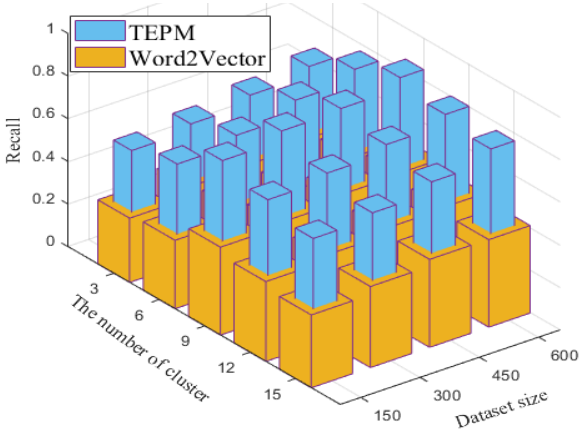
5.1 Evaluation environment and parameter settings

The proposed TEPM have collected a custom dataset which is gathered from major travel websites and utilised for training and evaluating our travel-related models, aiming to provide more accurate predictions and recommendations. This dataset comprises two components: a training set with 600 samples and a test set with 200 samples. These samples encompass a variety of travel destinations, packages, price ranges, and travel dates. They provide diversity and comprehensiveness for our research.

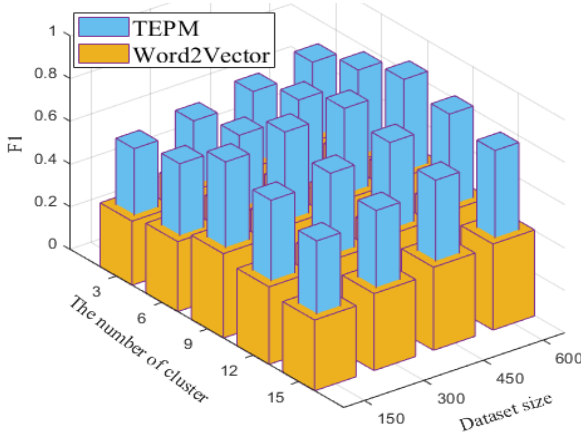
Figure 4 The proposed mechanism against Word2Vector, (a) precision (b) recall (c) F1 (see online version for colours)



(a)



(b)



(c)

5.2 Performance results

Figures 4(a), 4(b) and 4(c) compare the proposed TEPM against the Word2Vector in terms of precision, recall and F1. The dataset size varies from 150 to 600, and the number of clusters varies from 3 to 15. All the compared

mechanisms have the common trend that the precision, recall and F1 increase with the size of dataset. The reason is that a larger training dataset contributes to better model refinement. The precision, recall and F1 increase initially with the number of clusters when the number of clusters is below 9, but decrease when the number of clusters exceeds 9. This is because that the importance of selecting an appropriate number of clusters. An excessive number of clusters can lead to over-segmentation and reduce model effectiveness, emphasising the need for a balanced clustering strategy. The proposed TEPM outperforms the Word2Vector. The reason is that the proposed TEPM adopts FCC-BERT to propose the features from every advertisement and uses them to generate feature sentences, aiming to enhance the model performance.

Figure 5 Features of enrolment changes (see online version for colours)

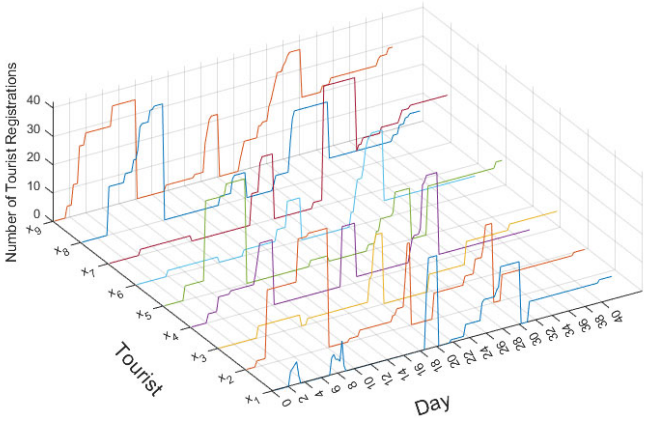


Figure 5 presents the features of enrolment changes in nine travel groups over a period of 24 weeks. As shown in Figure 5, it is evident that the enrolment fluctuations for each group exhibit diversity, indicating distinct trends and variations among different teams.

Figure 6 illustrates the changes in the number of tourist registrations, which have been normalised, for various travel groups within a 40-day period. The red line represents the daily registration predictions generated by the proposed TEMP for the travel groups in this cluster. It is evident that the predicted results closely resemble the actual registration trends, indicating the effectiveness of the algorithm. The reason is that the proposed TEMP adopts Two-tier clustering approach. The first-tier clusters are created by segregating the tour data according to the length of their registration days, while the second-tier clusters are formed by further partitioning the tour in the same first-tier cluster into several groups according to their similarity in terms of registration patterns. This method offers a more detailed representation of the similarities among the travel groups.

Figure 6 The prediction of TEPM (see online version for colours)

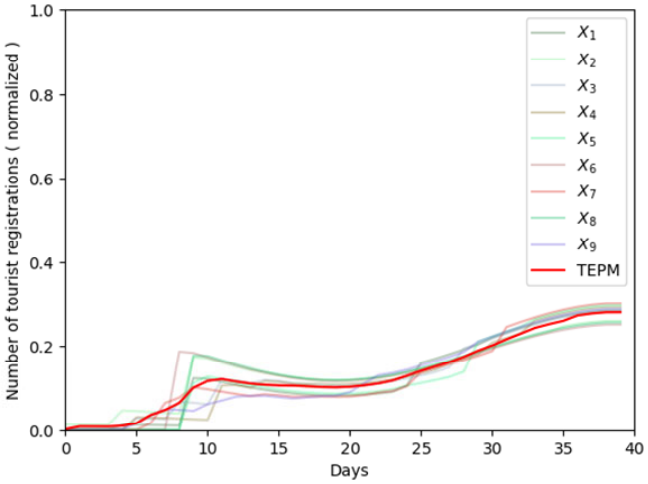
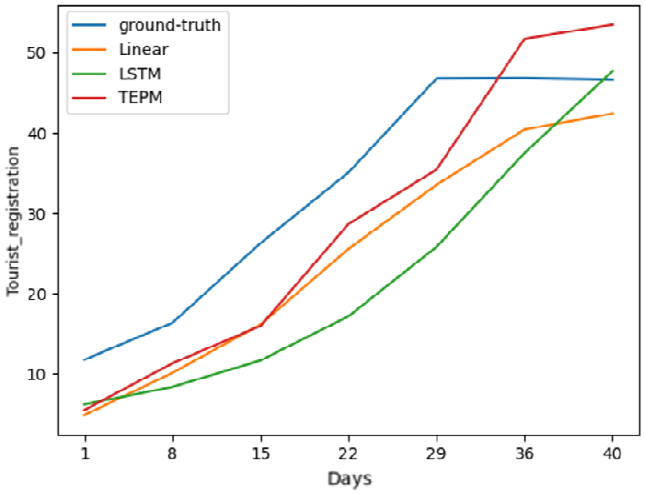
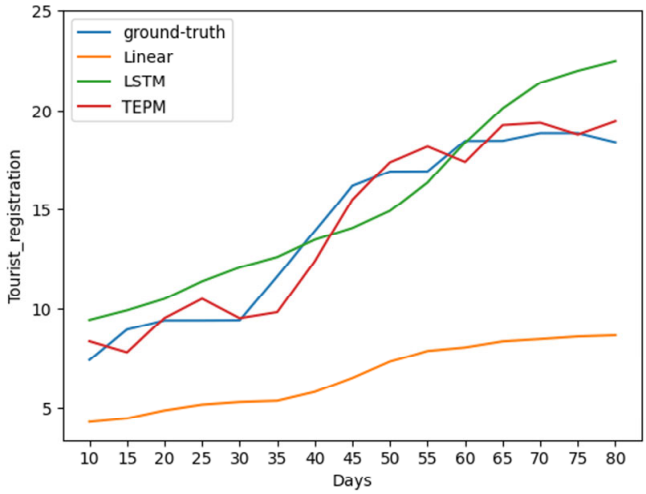


Figure 7 The prediction of tourist registrations, (a) 40 days (b) 80 days (c) 120 days (d) 160 days (see online version for colours)

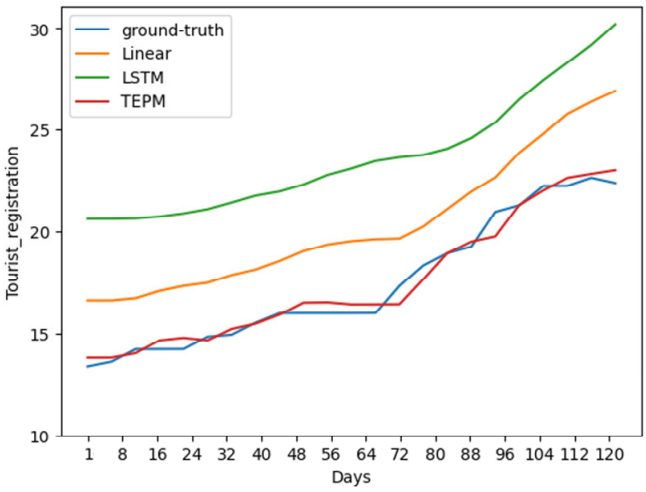


(a)

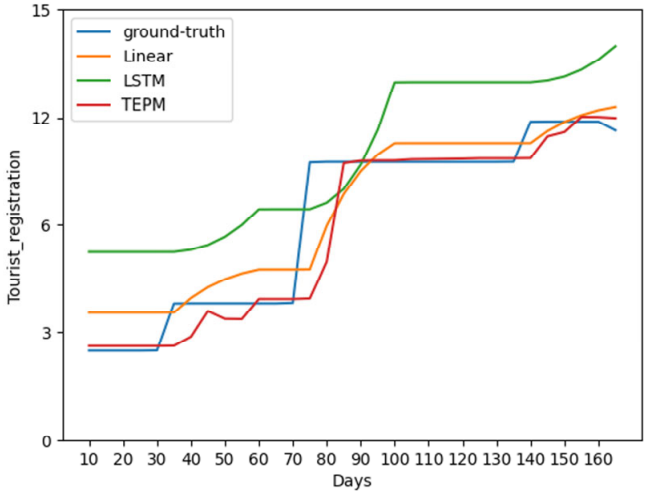


(b)

Figure 7 The prediction of tourist registrations, (a) 40 days (b) 80 days (c) 120 days (d) 160 days (continued) (see online version for colours)



(c)



(d)

Figures 7(a), 7(b), 7(c) and 7(d), a more comprehensive comparison, are made between the proposed TEPM, the Linear and LSTM models with regard to tourist registration. The data is grouped into four clusters based on the duration of 40 days, 80 days, 120 days, and 160 days. As shown in Figure 7, the proposed TEPM exhibits a trend that is closest to the ground truth. This proximity can be attributed to the approach employed, which first leverages BERT to extract relevant features from advertisements and associates them with predefined sub-clusters. Subsequently, an LSTM prediction model is applied to forecast future enrolment trends based on the registration patterns observed in the days leading up to enrolment. This two-tiered approach combines feature extraction, advertisement analysis, and prediction modelling, offering a more intelligent solution for travel group management and decision support.

6 Conclusions

This paper introduces the TEPM as a comprehensive approach to predict enrolment in tour groups. The proposed TEPM utilises available enrolment data for dynamic clustering to reflect changes in enrolment deadlines. Furthermore, the proposed TEPM conducts in-depth analysis of advertising content to identify key factors influencing traveller registrations. These factors are subsequently used as features, input into the FCC-BERT model for predicting the clustering of tour groups. Finally, LSTM-PMC is employed to forecast future enrolment trends based on prior registration data. Through these methods, the proposed TEPM provides a powerful tool for the travel industry to achieve more accurate and intelligent predictions of traveller enrolment behaviour and tour group management.

In future work, in addition to enrolment prediction, it is also possible to consider collecting feedback and satisfaction data from travellers to enhance the travel experience and the quality of tour groups.

Acknowledgements

This work was supported in part by 2022 Anhui University Humanities and Social Sciences Key Research Project No. 2022AH051069.

References

Abbasimehr, H., Shabani, M. and Yousefi, M. (2020) 'An optimized model using LSTM network for demand forecasting', *Computers & Industrial Engineering*, Vol. 143, p.106435.

Bi, J.W., Li, H. and Fan, Z.P. (2021) 'Tourism demand forecasting with time series imaging: a deep learning model', *Annals of Tourism Research*, Vol. 90, p.103255.

Cankurt, S. and Subasi, A. (2015) 'Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components', *Balkan Journal of Electrical and Computer Engineering*, Vol. 3, No. 1, pp.42–49.

Claveria, O., Monte, E. and Torra, S. (2015a) 'Combination forecasts of tourism demand with machine learning models', *Applied Economics Letters*, Vol. 23, No. 6, pp.428–431.

Claveria, O., Monte, E., and Torra, S. (2015b) 'Tourism demand forecasting with neural network models: different ways of treating information', *International Journal of Tourism Research*, Vol. 17, No. 5, pp.492–500.

He, K., Ji, L., Wu, C.W.D. and Tso, K.F.G. (2021) 'Using SARIMA-CNN-LSTM approach to forecast daily tourism demand', *Journal of Hospitality and Tourism Management*, Vol. 49, pp.25–33.

Kamel, N., Atiya, A.F., El Gayar, N. and El-Shishiny, H. (2008) 'Tourism demand foreacsting using machine learning methods', *ICGST International Journal on Artificial Intelligence and Machine Learning*, Vol. 8, pp.1–7.

Law, R., Li, G., Fong, D.K.C. and Han, X. (2019) 'Tourism demand forecasting: a deep learning approach', *Annals of Tourism Research*, Vol. 75, pp.410–423.

Peng, T., Chen, J., Wang, C. and Cao, Y. (2021) 'A forecast model of tourism demand driven by social network data', *IEEE Access*, Vol. 9, pp.109488–109496.

Sánchez-Medina, A.J. and Eleazar, C. (2020) 'Using machine learning and big data for efficient forecasting of hotel booking cancellations', *International Journal of Hospitality Management*, Vol. 89, p.102546.

Silva, E.S., Hassani, H., Heravi, S. and Huang, X. (2018) 'Forecasting tourism demand with denoised neural networks', *Annals of Tourism Research*, Vol. 74, pp.134–154.

Wang, H. and Liu, W. (2022) 'Forecasting tourism demand by a novel multi-factors fusion approach', *IEEE Access*, Vol. 10, pp.125972–125991.

Xia, H., Zhou, Y. and Zhang, Z. (2022) 'Auto insurance fraud identification based on a CNN-LSTM fusion deep learning model', *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 39, Nos. 1–2, pp.37–45.

Yao, Y. and Cao, Y. (2020) 'A Neural network enhanced hidden Markov model for tourism demand forecasting', *Applied Soft Computing*, Vol. 94, p.106465.

Zhang, B., Li, N., Shi, F. and Law, R. (2020a) 'A deep learning approach for daily tourist flow forecasting with consumer search data', *Asia Pacific Journal of Tourism Research*, Vol. 25, No. 3, pp.323–339.

Zhang, Y., Li, G., Muskat, B. and Law, R. (2020b) 'Tourism demand forecasting: a decomposed deep learning approach', *Journal of Travel Research*, Vol. 60, No. 5, pp.981–997.