Application of Classification and Word Embedding Techniques to Evaluate Tourists' Hotel-revisit Intention

Evripides Christodoulou[1], Andreas Gregoriades[1], Maria Pampaka[2] and Herodotos Herodotou[1]

[1]*Cyprus University of Technology, Limassol, Cyprus*

[2]*The University of Manchester, Manchester, U.K.*

Keywords: XGBoost, Topic Analysis, Word2Vec, Revisit Intention, Data Mining, Tourists' Reviews.

Abstract: Revisit intention is a key indicator for future business performance in the hospitality industry. This work focuses on the identification of patterns from user-generated data explaining the reasons why tourist may revisit a hotel they stayed at during their holidays and aims to identify differences among two classes of hotels (4-5 star and 2-3 star). The method utilises data from TripAdvisor retrieved using a scrapper application. Topic modelling is initially performed to identify the main themes discussed in each tourist review. Subsequently, reviews are labelled depending on whether they mention the intention of their author to revisit the hotel in the future using an ontology of revisit-intention generated using Word2Vec word embedding. The identified topics from the labelled reviews are utilised to train an Extreme Gradient Boosting model (XGBoost) to predict revisit intention, which is then used to identify topic-patterns in reviews that relate to revisit intention. The learned model achieved satisfactory performance and was used to identify the most influential topics related to revisit intention using an explainable machine learning technique to illustrate visually the rules embedded in the learned XGBoost model. The method is applied on reviews from tourists that visited Cyprus between 2009-2019. Results highlight that staff professionalism (e.g., politeness, smile) is critical for both classes of hotels; however, its effect is smaller on 2-3 start hotels where cleanliness has greater influence on revisiting.

1 INTRODUCTION

The rapid development of Web 2.0 applications has given users the opportunity for two-way communication and dramatically expanded tourism-related content on the Web (Sigala, 2016). This increase triggered the interest of many researchers in discovering patterns with regards to tourists preferences, behaviours, and actions (Pan et al., 2007). Online consumer opinions also referred as micro blogs communicated via social media such as Twitter, gained popularity as means for expressing peoples' views (Chamlertwat et al., 2012). Micro-blogs, also known as electronic word of mouth (eWOM), are a type of unstructured data relating to consumer opinions and emotions and can, therefore, affect an organisation's reputation. For this reason, eWOM analysis is linked with marketing activities. In tourism, there are two classes of tourist consumers: first-time consumers and returning consumers (Huang & Hsu, 2009). Destination marketers are keen to understand what drives tourists' intention to revisit, because the cost of retaining re-visitors is lower than that of attracting new visitors (Um et al., 2006); hence, the subject of tourist revisit has significant monetary benefits.

TripAdvisor is a popular platform for tourist related eWOM with numerous studies utilising such data to understand and measure customer satisfaction (Sotiriadis & van Zyl, 2013) or revisits. Most studies on tourists-revisit consider revisit intention as an extension of satisfaction that is derived from the experience during the initial visit. Most studies investigate the effect of eWOM on revisit intention using surveys (Huang & Hsu, 2009; Raza et al., 2012).

This research focuses on examining the concept of revisit intention by analysing eWOM generated by tourists after their experience with a hotel at a destination using natural language processing. Motivation for this work lies in the limited number of studies exploring the topics discussed in eWOM and their relationship to revisit intention.

Christodoulou, E., Gregoriades, A., Pampaka, M. and Herodotou, H.
Application of Classification and Word Embedding Techniques to Evaluate Tourists' Hotel-revisit Intention.
DOI: 10.5220/0010453502160223
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 216-223
ISBN: 978-989-758-509-8

The methodology employed includes the use of topic modelling to identify the main themes discussed in eWOM and word embeddings for the identification of reviews with high similarity with prespecified revisit keywords. The annotated data is used to train an XGBoost classification model that predicts revisit intention. The model's resulting patterns are made explicit through a popular machine leaning model interpretation technique, namely Shapley Additive explanations (SHAP).

The paper is organised as follows. The next section outlines the literature pertaining to revisit intention which is followed by a technical review of natural language processing techniques, covering topic modelling, word embedding, and XGBoost decision trees. Subsequent sections elaborate on the method followed and the results obtained. The paper concludes with the implications of the research and future directions.

2 LITERATURE REVIEW

This section briefly overviews the background on the area of tourist revisit intention which provides the context of this work

Revisit intention is the intention of a traveller to choose and visit more than once the same destination (Chang et al., 2019). The revisit intention procedure has been the focus of investigations for several years, trying to understand the patterns and variables that can lead to repeated purchase (Jang and Feng, 2007). According to destination choice theory, tourists select a destination that would satisfy their needs and get the planned outcome based on a range of criteria, such as economic factors, destination image, and destination environment (Stylos et al., 2017). During the decision process for choosing to visit for a second time the same destination, travellers consider the level of past satisfaction, needs satisfaction, attractiveness, and perceived quality as a metric of the difference between the expectations and the actual performance, and finally the overall perceived risk (Quintal & Polczynski, 2010).

Evidence suggests that many aspects of touristic experience as described in eWOM can affect revisit intention (Raza et al., 2012) while tourist accommodation experiences vary by hotel star rating and the ability of the hotels to provide different service quality that may lead to different parameters affecting the intention to return (Hu & Chen, 2016).

In this study, we use a bottom-up approach in identifying these factors utilizing eWOM produced by tourists. The main research questions are:

1. Which factors affect the intention of tourists to revisit a hotel based on the content of their online review?
2. What are the differences in the influencing factors between 4-5 and 2-3 hotels?

3 TECHNICAL OVERVIEW

Natural Language Processing refers to the field of research that focuses on the analysis/processing of human language using computers. It falls under the umbrella of artificial intelligence and uses machine learning, with the aim of finding patterns and information through raw data (Yilmaz, 2014). Models are trained using labelled or unlabelled data, with the former referred to as supervised and the latter as unsupervised learning. Regression and classification-based techniques such as XGBoost and Random Forest can be used to find patterns in textual data. Regression is used to predict values or ranges, such as the change in tourists' sentiment score from eWOM. Classification is used to find the category to which input data falls, for example finding whether a tourist intents to revisit a destination or not (Kim et al., 2020).

Natural language processing has many functions (Yilmaz, 2014); the most popular and relevant to this study are as follows:

1. Language Modelling, which focuses on modelling the input language so that it can understand and predict the possibility of the connections between words.
2. Information Retrieval, which is about finding relevant information linked to a keyword or keywords, for example "revisit intention".
3. Topic Modelling, which aims to discover the various topics discussed in a corpus of documents.

The remaining of this section elaborates on the main techniques in natural language processing that have been employed in this work. The emphasis is on topic modelling and word embedding methods as a mean to retrieve relevant information from the corpus and to label (i.e., specify the class they belong to) the reviews according to their connotation to revisit intention. The foundations of the XGBoost algorithm are also introduced.

3.1 Topic Modelling

Topic modelling is one of the basic functions of natural language processing and a very popular tool for extracting information from unstructured data. It belongs to the category of unsupervised data mining techniques employed to reveal and annotate documents with a key thematic information (Nikolenko et al., 2017). In general, it involves a statistical model aiming at finding topics that occur in a collection of documents. Two of the most popular techniques for topic analysis are the Latent Dirichlet Allocation and the Structural Topic Model (STM)

In this study, an STM topic model has been developed using the dataset (Roberts et al., 2014). Each topic represents a set of words occurring frequently together in a data set. Each topic is associated with different documents based on a probability distribution. In this paper, documents refer to online tourists' reviews.

3.2 Word Embedding

To process unstructured information in text, we need to transform this data into numerical format which can be processed. This transformation of raw text into a numeric vector format is known as word representation. Word embedding represents words as numerical vectors based on the contexts in which they appear and is a popular method for analysing text. This numerical word representation enables the mapping of words in a vocabulary to a point in a vector space. This technique is popular in capturing semantic relations among words usually learned from words co-occurrence information in large corpus. This technique is based on the distributional hypothesis stating that words occurring in similar contexts are semantically similar (Le & Mikolov, 2014). Word2vec is one of the most popular word embedding algorithm (Church 2017) that is based on artificial neural network, bag of words, and skip gram models (Church, 2017) to learn the mapping of words to a point in a vector space. Word2vec is a prediction-based method that can be implemented in two ways: as a continuous bag-of-words and as a skip-gram. Skip-gram is an unsupervised learning technique used to find the most related words to a given word, while the continuous bag-of-words attempts to predict a focus word given its context. The two key parameters for training word2vec are the number of the embedding dimensions, and the number of words before and after the target word, which is considered as its context. The advantages of word2vec compared to other word embedding algorithms is its ability to predict the probability of word's similarity to a given context and the ability to predict context similarity in a given word in addition to word similarity prediction (Church, 2017).

3.3 XGBoost

Extreme Gradient Boosting decision trees (XGBoost) is a newer version of the gradient boosting decision tree model (Chen & Guestrin, 2016) for classification and regression problems; it has been extensively used in academia and industry due to its excellent performance in machine learning tasks. XGBoost offers performance improvement over traditional boosting algorithms by addressing model's overfitting with regularization, and hence it produces generalizable models in timely manner (Gumus & Kiran, 2017). XGBoost combines multiple decision trees in a linear way, with each tree built based on the result of the previous tree (Zamani Joharestani et al., 2019). It is an ensemble method since it combines multiple classification and regression trees (CARTs), each composed of a number of nodes. Because XGBoost generates high performance on predictions, it has been used in research related to weather predictions, traffic accidents, etc. with the combination of SHAP for plotting models' insights (Parsa et al., 2020). Also, after validating its effectiveness in micro blogs, it has become a popular method and is often chosen for studies aiming to find patterns among eWOM text data for opinion mining (Wang et al., 2019).

4 METHODOLOGY

The workflow employed to answer our research questions is depicted in Figure 1. The process starts with the data extraction from TripAdvisor using a python scrapper for tourists who visited Cyprus between 2010 and 2019. The scraper saves in comma delimited format the travellers' username, rating of hotel, user helpful votes and contributions, date of stay and day of feedback, city of stay, hotel stars, country of origin, and the review text. A total of 65000 reviews was collected, all in English language, created by tourists coming from 27 countries and who stayed at 2 to 5-star hotels.

The first step in the process is the pre-processing of the data to remove irrelevant ASCII codes and the elimination of reviews with incomplete information (e.g., missing country of origin). Next step is the elimination of reviews from local visitors to focus on overseas tourists.

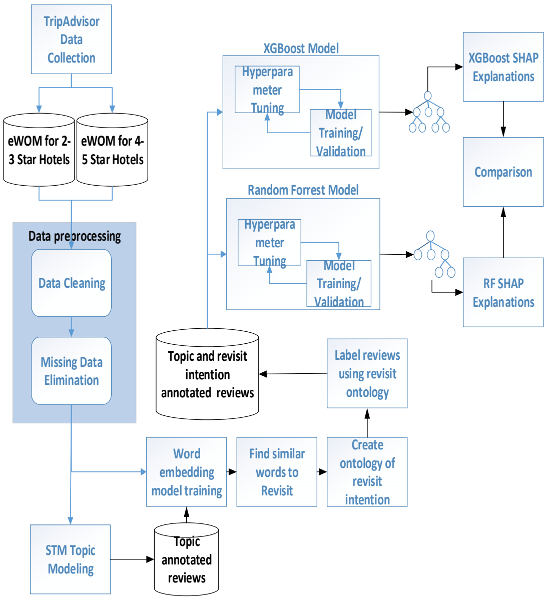
TripAdvisor Data Collection
eWOM for 2-3 Star Hotels
eWOM for 4-5 Star Hotels
Data preprocessing
Data Cleaning
Missing Data Elimination
STM Topic Modeling
Topic annotated reviews
Word embedding model training
Find similar words to Revisit
Create ontology of revisit intention
Label reviews using revisit ontology
Topic and revisit intention annotated reviews
XGBoost Model
Hyperparameter Tuning
Model Training/ Validation
Random Forrest Model
Hyperparameter Tuning
Model Training/ Validation
XGBoost SHAP Explanations
Comparison
RF SHAP Explanations

Figure 1: Overall method's flowchart.

The next step is the identification of the main themes presented in each review using the STM topic modelling approach. An illustration of this process can be found in Christodoulou *et al.*, (2020). During this process, the corpus underwent pre-processing to remove stop-words and irrelevant information, followed by tokenization (breaking sentences into word tokens), and stemming (converting words to their root form). During the learning of the STM model, an iterative process was followed examining different values for the number of topics (K) and inspecting the semantic coherence and exclusivity of the model at each iteration until a satisfactory model was produced (Roberts et al., 2014). Two different models were developed for this purpose, based on the feedback of tourists who stayed in two categories of hotels (2-3 and 4-5 stars).

The word2vec method is used to find the eWOM's author intention to revisit or not and was used as the class variable during the training of the XGBoost model. The features used for training the classification model were the topics that emerged from the STM results.

Labelling reviews based on the intention for revisit is a key process in this study. To make this possible, visit-related keywords were found from a learned word2vec model trained on the tourists' reviews corpus. A similar approach is employed to create ontologies for finding relevant documents from corpus (Church, 2017). The training of the word2vec models was performed using the pre-processed corpus using bigrams, trigrams, and four-grams phrases. Using the word2vec model and a similarity function of genism library, the 50 most similar words and phrases to revisit were identified, and from these, the 10 most correlated with each other were selected (Figure 2).

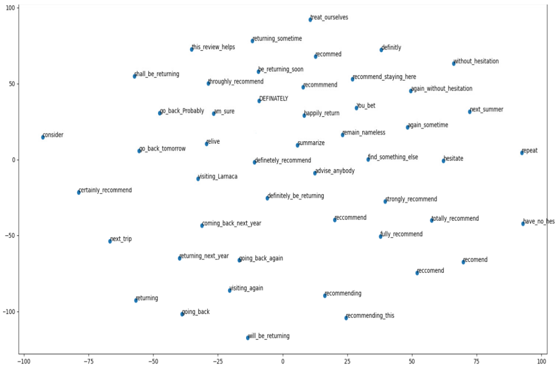

Figure 2: Most similar words to revisit using the learned word2vec model.

These words and their associations constituted a simple ontology based on which the corpus was searched, and reviews were labelled. The reviews were classified as positive if their content had a clear connotation of intent to revisit and neutral otherwise. Reviews with a clear reference to non-revisit intention were eliminated due to very low frequency/proportion. Reviews with no reference to revisit were considered as neutral. The intent to revisit was specified by searching the content of the reviews for text containing any of the words of phases in the prespecified ontology. To classify reviews that refer to revisit in a positive way rather than a negative, the rating of the review was also utilised. For instance, "I will revisit" versus "I will not revisit" both contain the keyword "revisit" but relating to different intentions. Hence, the rating of the review was used as an additional property to aid the correct labelling of the reviews, with high rating (>4) and occurrence of the key phases and words from the ontology denoting an intention to revisit and vice versa. A frequency distribution of the intention to revisit per year based on the compiled dataset is depicted in Figure 3 that also shows the imbalance between revisit and neutral classes.

The new dataset, that emerged from the labelled reviews and the topics associated with each review, was used to train a classification model using XGBoost to predict the topics that influence the intention to revisit by hotel-category. The topics' probability distributions per review were used as input features of the model. Given the data imbalance, and in order to maximize the efficiency of the model, the grid search of scikit toolkit was used to check and select the best hyperparameters for achieving the best Area Under the Curve (AUC) classification metric. This process was performed twice, for the two datasets relating to the class of hotel stars and the different topics per category.

To scrutinize further the learned XGBoost model and to enhance our confidence in the produced model results, an additional step was also included in our methodology that compared the XGBoost against another powerful tree-based classifier, namely Random Forest. This additional classifier was also developed using the same dataset and following the same steps as described earlier.

To improve the interpretability of the XGBoost model and aid the identification of patterns that have the most influence on the target variable, SHAP, a popular model interpretability technique was utilised. This enables the quantitative estimation of model interpretability and enables the visualization of black box algorithms outcome (Lundberg & Lee, 2017).

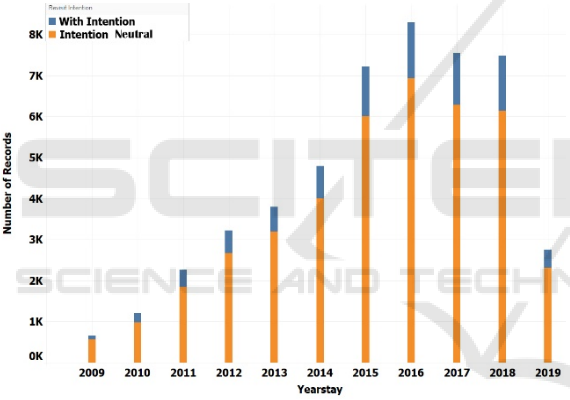

Figure 3: Distribution of revisit intention across the years in absolute numbers.

5 RESULTS

For the first stage of the process, the STM topic model was developed using the estimated K number of topics based on the model's coherence metrics. The K for the 2-3 and 4-5 stars hotels were 34 and 30 respectively. Topics were named using the most frequent words in each topic. The topics and their naming as identified are depicted in Table 1.

The word2vec model created using the reviews corpus was utilised to find the most similar words and phrases to revisit. Subsequently, similar words to the utput of that process were used again using the word2vec.

Table 1: STM topics table.

Num	Topic name 2 – 3 Stars	Topic name 4 - 5 Stars
1	Clubbing holidays	Renovated over years
2	Renovation	Poor
3	Good rooms	Special bar drinks
4	Dirty bathroom	Spa gym massage
5	Dirty-unprofessional staff	Smiling staff
6	Helpfull staff	Rude staff
7	Other guests	Efficient service
8	Insufficient all iincluusive	Best stayed
9	Good location	Well furnished room
10	Amazing staff	Extra costs for amenities (wifi,coffee machines)
11	Not well equipped room(eg.fridge)	Just ok
12	Good access (bus stations)	Amazing dinner (buffet,a la carte)
13	Comfortable room	Bad maintenance
14	Poor entertaiment and food	With transportations options
15	Limited breakfast options	Located close to beach
16	Pool area	Great for weddings
17	Value for money	Great for anniversaries
18	Close to beach with pool	Perfect location
19	Basic accomotation	Without complains agains reviews
20	Accept late arrials	Luxury
21	Good dinner	Great staff organisation
22	Low quality food and drinks	Room with seaview
23	Smelly room	Exceptional service
24	Good bar service	Dirty
25	Ideal for summer holidays (pool area)	Bad service
26	Bad customer service	With entertaiment activities
27	Old room	Great for families
28	Good entertaiment	Bad managment
29	Friendly staff	Centrally located
30	Basic apartment no luxuries	Great service staff
31	The best experiance	
32	Close to beach	
33	Bad facilities maintenance	
34	Clean good service	

This iterative process was continued until the words and phrases that were emerging had no connotation to revisit. A filtered subset of these words and phrases is depicted in table 2.

Table 2: Ontology Words retrieved from word2vec.

Keyword/Phase related to Revisit intention
Come back again
Be back
Be coming back again
Return again
Be going back
Be returning back
Consider returning
Most definitely return next year
Consider going back
Gladly return
Return sometime
Consider coming back
Come back next year
Consider staying
Come back
Most certainly come back
Return sometime

The relationships between these words formed the ontology based on which the reviews were classified. The XGBoost classification model that emerged from the compiled dataset (and after hyper parameter tuning) obtained a prediction accuracy of 84% and

AUC score of 90% for the reviews of tourists who visited 2-3 star hotels, and 89% prediction accuracy and AUC score of 90% for the 4-5 star hotels. Figures 4 and 5 show respectively the summary SHAP plots for the 2-3 and 4-5 star models with the topics' names depicted on the left y axis and the horizontal axis indicating the effect of each feature (topic) on the intention to revisit, with positive effect indicating higher log odds for revisit shown with a positive SHAP value on the x-axis. The weighting of the effect of each topic on the target variable is denoted by the hierarchy of the topics on the y axis with the most important topics appearing at the top. The colouring of the violin graphs indicate the intensity of the topics discussed in the reviews and the spread of each violin line denotes the frequency of each topics' state.

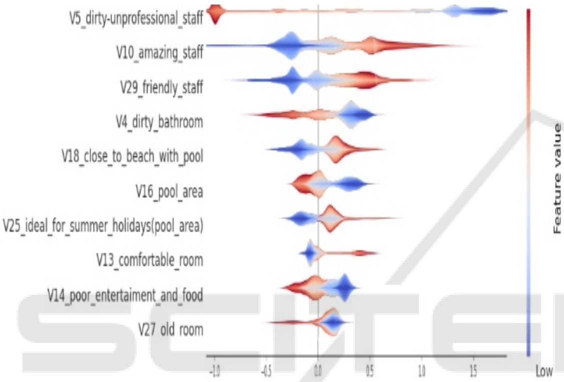

Figure 4: XGBoost summary for 2-3 star hotel reviews.

The model's most important topics that can shape the revisit intention are 10 from the total of 34 for the 2-3 star hotels. The most important topic refers to dirty space and unprofessional staff. When reviews refer to these, then the chance of re-visiting is reduced. The second and third most important topics mainly refer to staff in a positive way. When reviews talk about amazing and friendly staff, then the intention to revisit increases. The next most important topic has to do with the dirtiness of the rooms and especially bathrooms; when the tourist refers to such aspects, the revisit intention is reduced. The opposite happens when the review talks about the location of the hotel and its facilities as a summer destination.

The model for the 4-5 star hotels gave rise to the 10 most important topics that can determine the return visit. There are several similarities with the previous model for 2-3 star hotels. For example, the most important topic we find in the feedback text relates to the rudeness of the staff. When it appears, it reduces the chance of a tourist intending to visit again. The happy and smiley staff can lead to revisit intention, which can be reduced when the hospitality topic is reported to be cheap or poorly maintained or dirty.

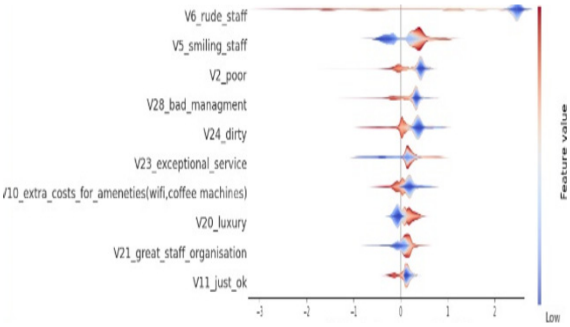

Figure 5: XGBoost summary for 4-5 star hotel reviews.

6 MODELS COMPARISON

To improve our confidences in the results from the XGBoost, a Random Forest (RF) model was also generated for each of the two hotel categories and compared against the results of the XGBoost models. The two RF models for each category of hotels yielded 82% prediction accuracy and 88% AUC score for the 2-3 star hotels, and 85% accuracy and 88% AUC score for the 4-5 star hotels. Both classifiers achieved comparable AUC performance, hence the purpose here was to investigate the inherent patterns of both models for possible discrepancies. Initially, the SHAP summary plots of the two RF models were generated and compared against those of XGBoost. Subsequently, the decision plots for both RF and XGBoost were compared as shown in Figure 6.

The decision plots in Figure 6 depict the first-order interactions in the model and hence display the cumulative effect of model's features and their interactions for one or more observations. In our case, we depict the average score of each feature from the dataset to explore how this, when used as input to the model, will generate an output. The x-axis shows the log odds of the revisit intention and in the y-axis the features in order of importance. The base value of the model is the average model output over the training dataset and is designated by the origin of the x-axis. The decision plot explains all predictions from the dataset using both main effects and interactions. As can be seen in Figure 6, the most influential topics to revisit intention that emerge from both models for 4-5 star hotels are very similar with the main topics affecting revisit intention that occur in both models, such as those that refer mainly to the staff and hotel cleanliness. For the development of the decision plots, we utilize the mean value for revisit intention and neutral cases, and use these two as exemplar cases to explain the critical mass of the dataset in an intuitive way. The decision plots highlight that the effect of the staff parameters on revisit intention are high with the negative effect (log odds) of poor staff attitude being higher than the positive effect of exceptional staff attitude. This abides with the asymmetry theory in service quality which states that the effect of a hotel attribute performance on guests' satisfaction is not linear. Similar results are obtained for the 2-3 start hotels.

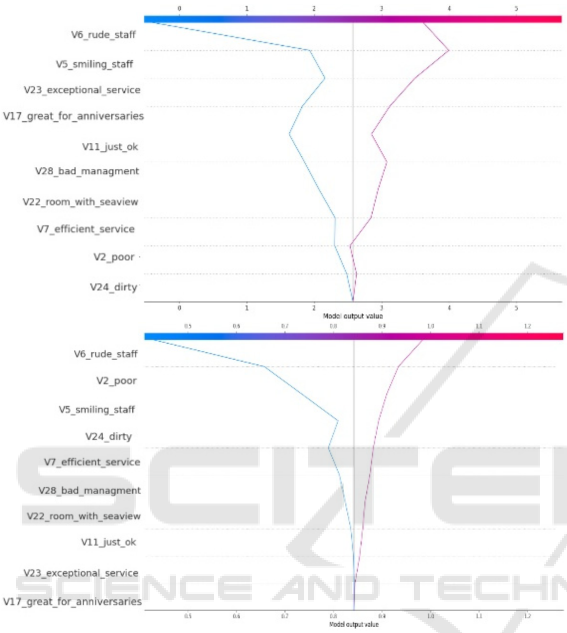

Figure 6: XGBoost (top) and RF (bottom). Decision plots for 4-5 stars hotels. The red line corresponds to revist intention cases and blue to neutral cases.

7 CONCLUSIONS

This study addressed the problem of tourists' revisit from a big data analytics perspective, given its importance to tourism industry. The methodology employed utilises topic modelling, word embedding, and contemporary classification algorithm such as XGBoost and Random Forest. Data was collected from TripAdvisor; topics were created using STM topic modelling and information retrieval using word2vec. Unlike other approaches, this method focuses on analysis of the raw text data posted on the internet in order to find the corresponding topics that correlate with the revisit intention instead of finding associations through predefined topics (Jang and Feng, 2007; Raza *et al.*, 2012).

The main results show that the variables that are associated with, and perhaps can influence, intention to return relate to aspects pertaining to the staff and service provided. The service and the emotions conveyed by the staff are important variables that shape the level of satisfaction (Rao, 2013) and the intention for a traveller to consider returning (Badarneh et al., 2001), thus these results are in agreement with literature. The next most important factors influencing the revisit intention are the cleanliness of the hotel, its facilities, and its location.

Similarities between the 2-3 and 4-5 star hotels are presented despite the difference in their stars rating. Even though the expectations of the tourists are different for each hotel class, the baseline factors remain similar (Shanka & Taylor, 2004).

The purpose of the data separation was to find different topics that would indicate different aspects influencing revisit intention in each hotel category. However, although the topics are different, the intention to return is associated with topics that are common to both cases and relate to staff professionalism and service. The main difference in the two classes of hotels is the level of cleanliness which has greater effect for the 2-3 star hotels compared to 4-5 hotels. Overall, the results highlight that staff professionalism is critical for both classes of hotels; however, its effect is smaller on 2-3 start hotels where cleanliness is also an important consideration for revisiting and has a greater effect.

A limitation of this work regards the labelling of reviews based on their revisit intention. The technique used in this study might miss out reviews that refer to revisit due to language semantics. To address this limitation a combination of embedding techniques will be applied in the future and the results will be compared with the proposed approach. Finally, future work will incorporate reviews in other languages.

REFERENCES

Jang, S., & Feng, R. (2007). Temporal destination revisit intention: The effects of novelty seeking and satisfaction. *Tourism Management*, *28*(2), 580–590.

Badarneh, M. B., Puad, A., & Som, M. (2001). Factors Influencing Tourists ' Revisit Behavioral Intentions and Loyalty. *Tourism Management*.

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. (2012). Discovering Consumer Insight from Twitter via Sentiment Analysis. *Journal of Universal Computer Science*, *18*(8), 973–992.

Chang, J.-R., Chen, M.-Y., Chen, L.-S., & Tseng, S.-C. (2019). Why Customers Don't Revisit in Tourism and Hospitality Industry? *IEEE Access*, *7*, 146588–146606.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Christodoulou, E., Gregoriades, A., Pampaka, M., & Herodotou, H. (2020). Combination of Topic Modelling and Decision Tree Classification for Tourist Destination Marketing. In *Lecture Notes in Business Information Processing* (pp. 95–108).

CHURCH, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162.

Gumus, M., & Kiran, M. S. (2017). Crude oil price forecasting using XGBoost. *2017 International Conference on Computer Science and Engineering (UBMK)*, 1100–1103.

Hu, Y.-H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management*, *36*(6), 929–944.

Huang, S., & Hsu, C. H. C. (2009). Effects of Travel Motivation, Past Experience, Perceived Constraint, and Attitude on Revisit Intention. *Journal of Travel Research*, *48*(1), 29–44.

Kim, J., Jang, S., Park, E., & Choi, S. (2020). Text classification using capsules. *Neurocomputing*.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *31st International Conference on Machine Learning, ICML 2014*.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102.

Pan, B., MacLaurin, T., & Crotts, J. C. (2007). Travel Blogs and the Implications for Destination Marketing. *Journal of Travel Research*, *46*(1), 35–45.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (Kouros). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident; Analysis and Prevention*, *136*, 105405.

Quintal, V. A., & Polczynski, A. (2010). Factors influencing tourists' revisit intentions. *Asia Pacific Journal of Marketing and Logistics*, *22*(4), 554–578.

Rao, P. S. (2013). Impact of Service Quality on Customer Satisfaction in Hotel Industry. *IOSR Journal Of Humanities And Social Science*, *18*(5), 39–44.

Raza, M., Siddiquei, A., Awan, H., & Bukhari, K. (2012). Relationship between service quality, perceived value, satisfaction and revisit intention in hotel industry. *Interdisciplinary Journal of Contemporary Research in Business*, 788–805.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Shanka, T., & Taylor, R. (2004). An Investigation into the Perceived Importance of Service and Facility Attributes to Hotel Satisfaction. *Journal of Quality Assurance in Hospitality & Tourism*, *4*(3–4), 119–134.

Sigala, M. (2016). Web 2.0 and customer involvement in new service development: A framework, cases and implications in tourism. In *Social Media in Travel, Tourism and Hospitality: Theory, Practice and Cases*.

Sotiriadis, M. D., & van Zyl, C. (2013). Electronic word-of-mouth and online reviews in tourism services: the use of twitter by tourists. *Electronic Commerce Research*, *13*(1), 103–124.

Stylos, N., Bellou, V., Andronikidis, A., & Vassiliadis, C. A. (2017). Linking the dots among destination images, place attachment, and revisit intentions: A study among British and Russian tourists. *Tourism Management*, *60*, 15–29.

Wang, S., Li, Z., Wang, Y., & Zhang, Q. (2019). Machine Learning Methods to Predict Social Media Disaster Rumor Refuters. *International Journal of Environmental Research and Public Health*, *16*(8), 1452.

Yilmaz, A. E. (2014). Natural Language Processing. *International Journal of Systems and Service-Oriented Engineering*, *4*(1), 68–83.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, *10*(7), 373.