DATA MINING TECHNIQUES FOR TOURIST REVIEW CLASSIFICATION

MUSTAPHA ABUBAKAR GIRO

A dissertation submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Computer Science

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

SEPTEMBER, 2019

DEDICATION

This dissertation is dedicated to my lovely parents Alhaji Abubakar Yusuf Giro and Hajiya Aishatu Alhassan for prayers and encouragement. It is also dedicated to my dearest wife Malama Salamatu Musa for her support and encouragement. And to my precious family for their continues prayers and support.

ACKNOWLEDGEMENT

In the Name of Allah, the Most Merciful, the Most Compassionate. All praises be to Allah, the lord of the worlds and may peace and blessing be upon the prophet Muhammad (S.A.W). I must acknowledge my endless thanks to Allah, the Ever-Magnificent, the Ever-Thankful, for His help and blessing. I am totally sure this work would never have become a reality without His guidance.

I wish to express my sincere appreciation to my supervisor, Dr. Alex Sim Tze Hiang, for his encouragement, guidance, motivation, extreme kindness and patience. Without his continued support and interest, this thesis would not have been the same as presented here. I would like to use this opportunity to say a warm thanks to the lecturers of the School of Computing that were involved directly or indirectly in assisting and guiding me to complete this dissertation.

I also would like to express my wholehearted thanks to all my family members for their generous support, love and prayers. Their views and tips were useful indeed. Last but not the least, deepest thanks go to all the people who took part in making this thesis a reality.

ABSTRACT

A large amount of information has been provided by the increasing volume of user generated content, through social networking services like reviews, comments and past experiences. Online review has become one of the most influential information sources for consumer decision-making. This information is freely accessible online and used to support tourist decision-making process. Despite several studies conducted on tourist online reviews, there have been limited studies exploring tourist reviews' ratings for 1 - 5 reviews star in predicting tourist response to an attraction. This study aims to predicting tourist ratings based on the tourist textual response (reviews) made on Petronas Twin Tower in Kuala Lumpur that is freely available on TripAdvisor. This is devised by building a predictive classification model that predicts the rating a tourist will possibly give. A qualitative approach is adopted where data miner tool was to collect tourist reviews from TripAdvisor; and the reviews dataset was preprocessed in Rapidminer to generate sentiment values which was fed to the models after some transformation. The sentiments gained/produced is utilized to compare which classification model gives the best prediction in terms of accuracy. The result showed that MLP prediction model returns a promising result in terms of accuracy over other techniques for predicting tourist response based on ratings (1-5) in which has 19% better accuracy than the other techniques tested. In conclusion, this study could contribute to the field of study by introducing a predictive model and could help destination marketers evaluate tourists' responses to a certain destination in advance, and could also potentially influence the final destination choice by improving marketing strategies accordingly. Destinations might use these analyses to predict the weaknesses or strengths of their image based on the analysis of tourists' reviews.

ABSTRAK

Sejumlah besar maklumat telah disediakan oleh bilangan pengguna yang banyak, melalui perkhidmatan rangkaian sosial seperti ulasan, komen dan pengalaman masa lalu. Ulasan melaui talian telah menjadi salah satu sumber maklumat yang paling berpengaruh untuk membuat keputusan pengguna. Maklumat ini boleh diakses secara dalam talian secara bebas dan digunakan untuk menyokong keputusan yang ingin dibuat oleh pelancong. Kesan ulasan dalam talian mengenai keputusan pelancong menarik beberapa kajian dewasa ini, hanya beberapa kajian mengkaji ulasan dalam talian berdasarkan gred dan belum ada yang langsung menganggap semua jenis gred dari ulasan 1 - 5 dalam meramalkan tarikan tertentu pelancong. Kajian ini bertujuan untuk meramalkan penilaian pelancong berdasarkan tindak balas tekstual pelancong (ulasan) yang dibuat di Menara Berkembar Petronas di Kuala Lumpur yang boleh didapati secara percuma di TripAdvisor. Kajian ini dibuat dengan membina model klasifikasi ramalan yang meramalkan gred pelancong yang akan diberi. Pendekatan kualitatif digunakan melalui perisian data miner dalam mengumpulkan ulasan pelancong dari TripAdvisor; dan data ulasan telah diproses menggunakan perisian Rapidminer untuk menghasilkan nilai sentimen yang diberi kepada model selepas beberapa perubahan. Teknik mining data digunakan untuk membandingkan dan mendapatkan prestasi ramalan optimum dari segi ketepatan MLP, SVM, DT, KNN, dan RF dalam MATLAB R2018a. Hasilnya menunjukkan bahawa model ramalan MLP mengembalikan hasil yang menjanjikan ketepatan berbanding teknik lain untuk meramalkan respon pelancong berdasarkan gred (1 - 5) di mana mempunyai ketepatan yang lebih baik 19% daripada teknik-teknik lain yang diuji. Sebagai kesimpulan, kajian ini dapat menyumbang kepada bidang pengajian dengan memperkenalkan model untuk meramalkan keutamaan pelancong berdasarkan penilaian untuk proses membuat keputusan para pengurus destinasi dan pengguna atas tarikan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Networks
API	-	Application Programming Interface
AI	-	Artificial Intelligence
ANOVA	-	Analysis of Variance
ADE	-	Adaptive Differential Evolution
AUC	-	Area Under Curve
BPNN	-	Back Propagation Neural Network
CFS	-	Correlation-based Feature Selection
CV	-	Cross Validation
DSS	-	Decision Support System
DT	-	Decision Trees
EUD	-	Euclidean Distance
EWUSC	-	Error Weighted Uncorrelated Shrunken Centroid
FP	-	False Positive
FN	-	False Negative
FNR	-	False Negative Rate
FPR	-	False Positive Rate
GDFM	-	Generalized Dynamic Factor Model
IBM	-	International Business Machine
IE	-	Information Extraction
KL	-	Kuala Lumpur
KNN	-	K-Nearest Neghbour
KDD	-	Knowledge Discovery in Databases
KDP	-	Knowledge Discovery Process
LM	-	Levenberg-Marquardt
MATLAB	-	MATrix LABoratory
MLP	-	Multilayer Perceptron

MI	-	Mutual Information
mRMR	-	Minimum Redundancy Maximum Relevance
MSE	-	Mean Squared Error
NPL	-	Natural Processing Language
NP-Hard	-	Non-deterministic Polynomial-time Hard
PCA	-	Principal Component Analysis
RBF	-	Radial Basis Function
RE	-	Relation Extraction
RF	-	Random Forest
RMSE	-	Root Mean Squared Error
ROC	-	Receiver Operating Characteristic
SNS	-	Social Networking Sites
SVM	-	Support Vector Machine
TNR	-	True Negative Rate
TPR	-	True Positive Rate
TN	-	True Negative
TP	-	True Positive
UGC	-	User-Generated Content

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Online reviews by other tourist or visitors are used to provide assistance for new tourists to decide on the choice of where to visit (such as tourist center, hotels, business places and so on). Reviews are used not only in tourist attraction but widely used in e-commerce for products recommendation to other customers where previous customers expressed their satisfaction or dissatisfaction on a particular product based on their experience with it (Pantano *et al.*, 2017). In most cases, the reviews/comments turn to be accurate or correct. This is one of the motivations behind tourist attraction prediction using reviews.

The rapid spreading of User Generated Content by means of social media services has generated large quantities of data accessible to people. Visitors routinely access these data to enable their essential decision-making process. The data is openly accessible on the web in the form of tourist reviews or online reviews (Pantano *et al.*, 2017). The present advances in electronic media innovations and environment, together with the intelligent application software for automated advertising in the data society age (Buhalis and Law, 2008), are keys to the increase in internet users thereby increase the numbers of reviewers that makes online reviews from thousands to millions.

1.2 Problem Background

Reviewers' comments are used as guide in e-commerce for product recommendations. The use of reviews in decision making is not limited to e-commerce but also applied to generation of tourist recommendations and predictions. The traditional ways of making reviews are normally on e-commerce websites for products

and tourist websites for tourist recommendation. The present advances in digital media and the use of intuitive software applications have motivated advances in advertising in the era of information. Online marketing in particular with its basic requirement of user account creation has changed the way data is gotten to and shared (Pantano *et al.*, 2017).

Digital marketers understand that to successfully attract and influence the interest of SNS users, they need to increase the utility of social networks by offering value added services (Lu and Stepchenkova, 2012). Therefore, SNSs are now increasing their capabilities by offering a various portfolio of build-in applications (apps) to meet social media users' needs for better experiences; like, customized topic-specific virtual spaces to better support User-Generated Content (UGC) (e.g. Facebook apps, YouTube), including reviews, comments on past experiences and recommendations for future visits (Riordan *et al.*, 2016). As researchers note, online reviews based on SNS users' profiles and established preferences are basic to formulating future preferences and affecting consumer purchases. (Baka, 2016b). It was assumed that people's behavior towards an attraction/destination is basically determined by people's comments, beliefs, feelings, recommendation, and prior encounter (Pantano *et al.*, 2017).

In fact, the more the product online review features available to consumers, the higher the likelihood for sales of related items within the product category (Chevalier and Mayzlin, 2006). Also, in a travel and tourism context, tourists' recommendations via tripadvisor, Yelp, expedia etc. influence other travelers' decisions about many different aspects of their trips, e.g. selection of a tourist destination, accommodation and attractions to visit (Bilgihan *et al.*, 2016). besides the fact that some studies (Fotis *et al.*, 2011; Lin *et al.*, 2018; Pantano *et al.*, 2017) have shown that many reviews are fake, or too positive or negative, consumers perceive online reviews as more trustworthy than content provided by official destination websites (Afzaal *et al.*, 2019). Drawing on a huge amount of UGC, marketers make systematic efforts to exploit as much open data as possible to support digital marketing effectiveness. These efforts could potentially improve online sales and the profitability of e-travel services (e.g., accommodation, transportation, restaurants, entertainment, sightseeing and tourism

destination information) (Pantano *et al.*, 2017).

Most researches in User-Generated Content (UGC) and most online reviews have underlined the importance of analyzing ratings to increase the likelihood of travelers' having enjoyable trips (Pantano *et al.*, 2017). However, few studies investigate the effect of reviews on SNS tourists' future decisions (Amaral *et al.*, 2014; Berger *et al.*, 2010; Chintagunta *et al.*, 2010; Kim *et al.*, 2016; Pantano *et al.*, 2017). These studies focus on the readability, credibility and helpfulness of online reviews, however they do not explore the extent to which recommendations maybe perceived as useful to other travelers willing to travel to the same destinations (Filieri, 2016). Online marketing in particular with its basic requirement of user account creation has changed the way data is gotten to and shared (Melián-González *et al.*, 2013; Pantano *et al.*, 2017; Torres *et al.*, 2015).

However, machine learning and data mining techniques could be used to analyse data generated from tripadvisor.com in order to recognize useful patterns, such as the classification of new tourist responses. The process of developing a model that will predict the class of new unlabelled data is called a classification task which in machine learning techniques is referred to as supervised learning (Ando and Zhang, 2005; Collobert and Weston, 2008; Kotsiantis *et al.*, 2007). The data to be used for that model construction consist of a set of features that will be used to train the model.

From a theoretical point of view, this study draws attention to the potential of using online reviews that are freely available to influence tourist's attitudes and behaviors. consequently, this research proposes a computational tool that contributes to the effective positioning of hospitality organizations and tourist destinations/attractions in tourism management.

To apply these methods in feature selection task, mapping concepts into the field of feature selection is required. Figure 1.1 below describes the problem and issues of previous work in the prediction process.

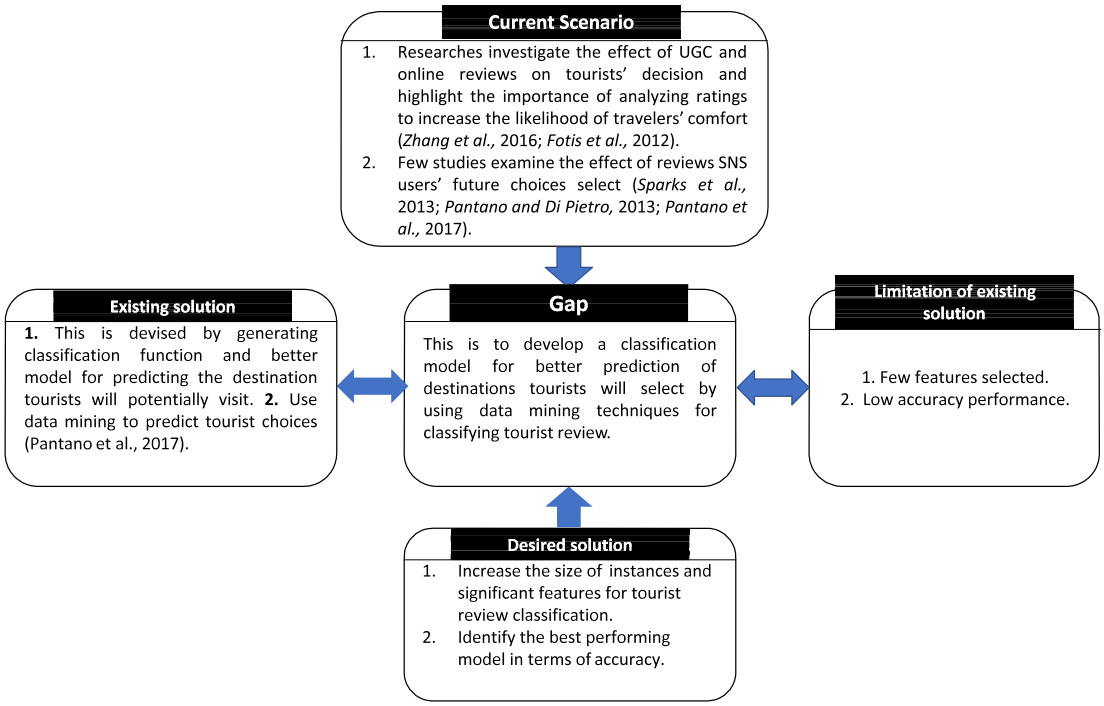

Figure 1.1 Problem of Previous work in the Prediction Process

1.3 Problem Statement

Despite recent researches conducted in tourist reviews to obtain reasonable performance in terms of accuracy for predicting tourist response to a tourist attraction/destination based on ratings (Fang *et al.*, 2016; Jacobsen and Munar, 2012; Lee *et al.*, 2018a; Moro *et al.*, 2017a; Pabel and Pearce, 2015; Pantano *et al.*, 2017), however, none of these used a sentiment analyzer operator, an aylien service available in Rapid miner that returns review's subjectivity and polarity values to be fed into the multilayer perceptron algorithm. And also, none used all the traveler ratings (from 1 -5 stars) to predict the tourist response to a certain attraction. This research seeks to include all traveler ratings and improves prediction performance in terms of accuracy.

Feature selection methods are effective in removing redundant and irrelevant features, improving learning algorithm's prediction performance and are also needed when the number of training examples is too tittle, or when there is too much data that can be processed efficiently by the machine learning algorithms (Senliol *et al.*, 2008).

1.4 Research Questions

To address the underlying issues, the main research question for this study is "How can a prediction performance of tourist response to an attraction be improved?" To support the main research question three (3) sub questions were formulated as follows: -

(a) How to identify significant features of tourist reviews on Tripadvisor for all ratings?

(b) How to predict tourist response and demand, based on ratings, towards a location using classification techniques for tourist review?

(c) How to compare the MLPs' prediction accuracy with other existing techniques, SVM, DT, RF, and KNN?

1.5 Goal of the Study

The aim of this research is to determine best tourist decision prediction model using review's rating that returns the best accuracy.

1.6 Research Objectives

The objectives of the research are:

(a) To identify significant features of tourist reviews on Tripadvisor for all 1 - 5 star ratings.

(b) To predict tourist response and demand, based on ratings, towards a location using classification techniques for tourist review.

(c) To compare the prediction accuracy of MLP, SVM, DT, RF, and KNN classification techniques.

1.7 Scope of the Study

To achieve the above-mentioned objectives, this research is limited to the following:

(a) Multilayer Perceptron is used to predict tourist response based on ratings.

(b) Dataset used in this study is obtained from Tripadvisor's tourist reviews made on Petronas Twin Towers in Kuala Lumpur, Malaysia for two (2) years (2017 - 2018).

(c) This study used an Aylien technique in RapidMiner for feature selection.

(d) MATLAB R2018a academic release is used for classification model simulation.

1.8 Significance of the Study

This research is significant from both theoretical and practical perspectives. The rationale and motivation for this research are:

(a) Considering a specific attraction reviewed in TripAdvisor, this research is essential in identifying the trend of consumers' appreciation of a certain tourist destination/attraction. Therefore, tourism managers can consider adopting tourist reviews analysis for better predictions about the attractiveness of a certain destination.

(b) This research helps destination marketers in advance to evaluate tourists' responses to a certain destination, and accordingly can potentially influence potential destination options by improving marketing strategies. Moreover, tourist destinations managers can use this analysis to predict the weaknesses or strengths of their image based on the analysis of tourists' reviews, which is easily accessible online.

(c) This study draws attention to the large potential of using tourist online reviews sources to influence tourists' attitude and behavior. In practice, we propose a computational tool that can contribute greatly to the effective positioning

REFERENCES

Abaker, I., Hashem, T., Yaqoob, I., Badrul, N., Mokhtar, S., Gani, A. and Ullah, S. (2015). The rise of "big data" on cloud computing : Review and open research issues. *Information Systems*. 47, 98–115. ISSN 0306-4379. doi:10.1016/j.is.2014.07.006. Retrievable at http://dx.doi.org/10.1016/j.is.2014.07.006 .

Afzaal, M., Usman, M., Fong, A. C. and Fong, S. (2019). Multiaspect-based opinion classification model for tourist reviews. *Expert Systems*. 36(August 2018), 1–24. ISSN 14680394. doi:10.1111/exsy.12371.

Aggarwal, C. (2011). *Social Network Data Analytics*. Boston, MA: Springer US. ISBN 978-1-4419-8461-6. doi:10.1007/978-1-4419-8462-3. Retrievable at http://medcontent.metapress.com/index/A65RM03P4874243N.http://link.springer.com/10.1007/978-1-4419-8462-3 .

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Alhelalat, J. A., Habiballah, M. A. and Twaissi, N. M. (2017). The impact of personal and functional aspects of restaurant employee service behaviour on customer satisfaction. *International Journal of Hospitality Management*. 66, 46–53. ISSN 02784319. doi:10.1016/j.ijhm.2017.07.001. Retrievable at http://dx.doi.org/10.1016/j.ijhm.2017.07.001 .

Alkalbani, A. M., Gadhvi, L., Patel, B., Hussain, F. K., Ghamry, A. M. and Hussain, O. K. (2017). Analysing cloud services reviews using opining mining. *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, 1124–1129. ISSN 1550445X. doi:10.1109/AINA.2017.173.

Alshamlan, H. M., Badr, G. H. and Alohali, Y. A. (2016). Abc-svm: artificial bee colony and svm method for microarray gene selection and multi class cancer classification. *Int. J. Mach. Learn. Comput*. 6(3), 184.

Amaral, F., Tiago, T. and Tiago, F. (2014). User-generated content: tourists' profiles on Tripadvisor. *International Journal of Strategic Innovative Marketing*. 1(3), 137–145.

Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from

multiple tasks and unlabeled data. *Journal of Machine Learning Research*. 6(Nov), 1817–1853.

Baka, V. (2016a). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. *Tourism Management*. 53, 148–162.

Baka, V. (2016b). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. *Tourism Management*. 53, 148–162. ISSN 02615177. doi:10.1016/j.tourman.2015.09.004.

Balbi, S., Misuraca, M. and Scepi, G. (2018). Combining different evaluation systems on social media for measuring user satisfaction. *Information Processing & Management*. 54(4), 674–685.

Barbier, G. and Liu, H. (2011). Data mining in social media. In *Social network data analytics*. (pp. 327–352). Springer.

Basiri, M. E., Ghasem-Aghaee, N. and Reza, A. (2017). Lexicon-based sentiment analysis in Persian. *Curr. Futur. Dev. Artif. Intell.*, 154.

Bellman, R. E. (2015). *Adaptive control processes: a guided tour*. vol. 2045. Princeton university press.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*. 5(Sep), 1089–1105.

Berger, J., Sorensen, A. T. and Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*. 29(5), 815–827.

Bhattacharya, G., Ghosh, K. and Chowdhury, A. S. (2017). Granger Causality Driven AHP for Feature Weighted kNN. *Pattern Recognition*. 66(May 2016), 425–436. ISSN 00313203. doi:10.1016/j.patcog.2017.01.018. Retrievable at https://linkinghub.elsevier.com/retrieve/pii/S0031320317300195 .

Bilgihan, A., Barreda, A., Okumus, F. and Nusair, K. (2016). Consumer perception of knowledge-sharing in travel-related Online Social Networks. *Tourism Management*. 52, 287–296. ISSN 0261-5177. doi:10.1016/j.tourman.2015.07.002. Retrievable at http://dx.doi.org/10.1016/j.tourman.2015.07.002 .

Breiman, L. (2001). Random forests. *Machine learning*. 45(1), 5–32.

Buchler, M., Franzini, G., Franzini, E. and Eckart, T. (2016). Mining and analysing one billion requests to linguistic services. In *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*. ISBN 9781467390040, 3230–3239. doi: 10.1109/BigData.2016.7840979.

Buhalis, D. and Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet - The state of eTourism research. *Tourism management*. 29(4), 609–623.

Cambria, E., Schuller, B., Xia, Y. and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*. 28(2), 15–21.

Chen, M.-S., Han, J. and Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*. 8(6), 866–883.

Chevalier, J. A. and Mayzlin, D. (2006). The Effect of Word of Mouth on Sales : Online Book Reviews. *JOURNAL OF MARKETING RESEARCH*. XLIII(August), 345–354.

Chintagunta, P. K., Gopinath, S. and Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*. 29(5), 944–957.

Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*. 34(2), 272–284. ISSN 02684012. doi:10.1016/j.ijinfomgt.2014.01.001.

Claveria, O., Monte, E. and Torra, S. (2015). Common trends in international tourism demand: Are they useful to improve tourism predictions? *Tourism Management Perspectives*. 16, 116–122. ISSN 22119736. doi:10.1016/j.tmp.2015.07.013. Retrievable at http://dx.doi.org/10.1016/j.tmp.2015.07.013 .

Clavería, Ó., Monte Moreno, E. and Torra Porras, S. (2015). Effects of removing the trend and the seasonal component on the forecasting performance of artificial neural network techniques. *AQR–Working Papers, 2015, AQR15/03*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 160–167.

Cox, C., Burgess, S., Sellitto, C. and Buultjens, J. (2008). Consumer-generated web-

based tourism marketing. *Cooperative Research Centre for Sustainable Tourism, Australia*.

Cristianini, N., Shawe-Taylor, J. *et al.* (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

De León-Aldaco, S. E., Calleja, H. and Alquicira, J. A. (2015). Metaheuristic optimization methods applied to power converters: A review. *IEEE Transactions on Power Electronics*. 30(12), 6791–6803.

Devaraja, R. R. (2018). *Project AISRA (Autonomous interface for stimulus robust application)*. Ph.D. Thesis. Kaunas University of Technology.

Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*. 22, 457–479. ISSN 10769757. doi:10.1613/jair.1523.

Fang, B., Ye, Q., Kucukusta, D. and Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*. 52, 498–506. ISSN 02615177. doi:10.1016/j.tourman.2015.07.018.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*. 17(3), 37–37.

Filieri, R. (2016). Annals of Tourism Research What makes an online consumer review trustworthy ? *Annals of Tourism Research*. 58, 46–64. ISSN 0160-7383. doi:10.1016/j.annals.2015.12.019. Retrievable at http://dx.doi.org/10.1016/j.annals.2015.12.019 .

Fister, I., Fister Jr, I., Yang, X.-S. and Brest, J. (2013). A comprehensive review of firefly algorithms. *Swarm and Evolutionary Computation*. 13, 34–46.

Fotis, J., Buhalis, D. and Rossides, N. (2011). Social media impact on holiday travel planning: The case of the Russian and the FSU markets. *International Journal of Online Marketing (IJOM)*. 1(4), 1–19.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*. 35(2), 137–144. ISSN 02684012. doi:10.1016/j.ijinfomgt.2014.10.007. Retrievable at http://linkinghub.elsevier.com/retrieve/pii/S0268401214001066 .

Ganesan, K., Zhai, C. and Han, J. (2010). Opinosis : A Graph-Based Approach

to Abstractive Summarization of Highly Redundant Opinions. August. 340–348. doi:http://portal.acm.org/citation.cfm?id=1873781.1873820.

García, S., Luengo, J. and Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

Ghosh, R. and Lerman, K. (2010). Predicting influential users in online social networks. *arXiv preprint arXiv:1005.4882*.

Glover, F. W. and Kochenberger, G. A. (2006). *Handbook of metaheuristics*. vol. 57. Springer Science & Business Media.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*. 3(Mar), 1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. 11(1), 10–18.

Han, J., Pei, J. and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hanna, V., Backhouse, C. and Burns, N. D. (2004). Linking employee behaviour to external customer satisfaction using quality function deployment. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*. 218(9), 1167–1177.

He, W., Zha, S. and Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*. 33(3), 464–472.

Hira, Z. M. and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*. 2015.

Jacobsen, J. K. S. and Munar, A. M. (2012). Tourist information search and destination choice in a digital age. *Tourism Management Perspectives*. 1, 39–47. ISSN 22119736. doi:10.1016/j.tmp.2011.12.005.

Karahoca, A., Karahoca, D. and Anver, M. (2012). Survey of Data Mining and Applications (Review from 1996 to Now). In *Data Mining Applications in Engineering and Medicine*. InTech. doi:10.5772/48803.

Kim, M. J., Lee, C.-K. and Bonn, M. (2016). The effect of social capital and altruism

on seniors' revisit intention to social network sites for tourism-related purposes. *Tourism Management*. 53, 96–107.

Kohavi, R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, vol. 14. Montreal, Canada, 1137–1145.

Kotsiantis, S. B., Zaharakis, I. and Pintelas, P. (2007). *Supervised machine learning: A review of classification techniques*. vol. 160.

Kurgan, L. A. and Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*. 21(1), 1–24.

Kusiak, A. and Li, W. (2011). The prediction and diagnosis of wind turbine faults. *Renewable Energy*. 36(1), 16–23.

Larose, D. T. and Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Larose, D. T. and Larose, D. T. (2006). *Data mining methods and models*. vol. 12. Wiley Online Library.

Lee, P.-J., Hu, Y.-H. and Lu, K.-T. (2018a). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics*. ISSN 07365853. doi:10.1016/j.tele.2018.01.001.

Lee, P.-J., Hu, Y.-H. and Lu, K.-T. (2018b). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics*. 35(2), 436–445.

Li, G., Law, R., Vu, H. Q., Rong, J. and Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism management*. 46, 311–321.

Li, S., Chen, T., Wang, L. and Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*. 68, 116–126.

Li, X., Pan, B., Law, R. and Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism management*. 59, 57–66.

Liao, S.-H., Chu, P.-H. and Hsiao, P.-Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert systems with applications*.

39(12), 11303–11311.

Liao, S.-H. and Wen, C.-H. (2007). Artificial neural networks classification and clustering of methodologies and applications–literature analysis from 1995 to 2005. *Expert Systems with Applications*. 32(1), 1–11.

Lin, P. M., Tung, V. W. S., Qiu Zhang, H. and Gu, Q. (2018). Tourist experience on memorable hospitality services. *Journal of China Tourism Research*. 14(2), 123–145.

Litvak, M. and Last, M. (2008). Graph-based keyword extraction for single-document summarization. August. ISBN 9781905593514, 17. doi:10.3115/1613172.1613178. Retrievable at http://portal.acm.org/citation.cfm?doid=1613172.1613178 .

Lu, W. and Stepchenkova, S. (2012). Ecotourism experiences reported online: Classification of satisfaction attributes. *Tourism Management*. 33(3), 702–712. ISSN 02615177. doi:10.1016/j.tourman.2011.08.003.

Melián-González, S., Bulchand-Gidumal, J. and González López-Valcárcel, B. (2013). Online customer reviews of hotels: As participation increases, better evaluation is obtained. *Cornell Hospitality Quarterly*. 54(3), 274–283.

Mitiche, A. and Lebidoff, M. (2001). Pattern classification by a condensed neural network. *Neural networks*. 14(4-5), 575–580.

Moro, S., Rita, P. and Coelho, J. (2017a). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. *Tourism Management Perspectives*. 23, 41–52. ISSN 22119736. doi:10.1016/j.tmp.2017.04.003.

Moro, S., Rita, P. and Coelho, J. (2017b). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. *Tourism management perspectives*. 23, 41–52.

Munar, A. M., Kr, J. and Jacobsen, S. (2013). Trust and Involvement in Tourism Social Media and Web-Based Travel Information Sources. *Scandinavian Journal of Hospitality and Tourism*. 2250. doi:10.1080/15022250.2013.764511.

Murray-Rust, P. (2008). Open Data in Science. *Serials Review*. 34(1), 52–64. ISSN 00987913. doi:10.1016/j.serrev.2008.01.001. Retrievable at `http://linkinghub.elsevier.com/retrieve/pii/S009879130800004X` .

Ngai, E. W., Xiu, L. and Chau, D. C. (2009). Application of data mining techniques

in customer relationship management: A literature review and classification. *Expert systems with applications*. 36(2), 2592–2602.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Oracle (2013). Big Data Analytics (Advanced analytics in oracle database). (March), 1–13.

Özköse, H., ArÄś, E. S. and Gencer, C. (2015). Yesterday, Today and Tomorrow of Big Data. *Procedia - Social and Behavioral Sciences*. 195, 1042–1050. ISSN 18770428. doi:10.1016/j.sbspro.2015.06.147.

Pabel, A. and Pearce, P. L. (2015). Highlighting the benefits of tourism humour: The views of tourists. *Tourism Management Perspectives*. 16, 357–364. ISSN 22119736. doi:10.1016/j.tmp.2015.10.002.

Pal, M. and Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*. 48(5), 2297–2307.

Pantano, E. and Di Blasi, G. (2015). Big data analysis solutions for supporting tourist's decision making. In *Proceedings of the international conference on tourism (ICOT2015)*. International Association for Tourism Policy London, UK, 440e452.

Pantano, E., Priporas, C.-V. and Stylos, N. (2017). 'You will like it!' using open data to predict tourists' response to a tourist attraction. *Tourism Management*. 60, 430–438. ISSN 02615177. doi:10.1016/j.tourman.2016.12.020. Retrievable at `https://` linkinghub.elsevier.com/retrieve/pii/S0261517716302680 .

Petty, G. P. (2012). *Active Learning*. ISBN 9781608457250.

Phillips, P., Zigan, K., Silva, M. M. S. and Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management*. 50, 130–141.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Rajaraman, A. and Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Renuka, D. K., Hamsapriya, T., Chakkaravarthi, M. R. and Surya, P. L. (2011). Spam classification based on supervised learning using machine learning techniques. In *2011 International Conference on Process Automation, Control and Computing*. IEEE, 1–7.

Riordan, S. O., Feller, J. and Nagle, T. (2016). A categorisation framework for a feature-level analysis of social network sites social network sites. *Journal of Decision Systems*. 0125(June). doi:10.1080/12460125.2016.1187548.

Rojas, R. (2013). *Neural networks: a systematic introduction*. Springer Science & Business Media.

Saeys, Y., Inza, I. and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*. 23(19), 2507–2517. ISSN 13674803. doi:10.1093/bioinformatics/btm344.

Sayadi, A. R., Fathianpour, N. and Mousavi, A. A. (2011). Open pit optimization in 3D using a new artificial neural network. *Archives of Mining Sciences*. 56(3), 389–403.

Senliol, B., Gulgezen, G., Yu, L. and Cataltepe, Z. (2008). Fast Correlation Based Filter (FCBF) with a different search strategy. *2008 23rd International Symposium on Computer and Information Sciences, ISCIS 2008*. doi:10.1109/ISCIS.2008.4717949.

Sivarajah, U., Kamal, M. M., Irani, Z. and Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*. 70, 263–286. ISSN 01482963. doi:10.1016/j.jbusres.2016.08.001.

Strahilevitzt, L. J. (2005). A Social Networks Theory of Privacy Author (s): Lior Jacob Strahilevitz Source : The University of Chicago Law Review , Vol . 72 , No . 3 (Summer , 2005), pp . 919-988 Published by : The University of Chicago Law Review Stable URL : http://www.jstor.o. 72(3), 919–988.

Sutha, K. and Tamilselvi, J. J. (2015). A review of feature selection algorithms for data mining techniques. *International Journal on Computer Science and Engineering*. 7(6), 63.

Tang, H., Tan, S. and Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*. 36(7), 10760–10773.

Thomas, J., McNaught, J. and Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*. 2(1), 1–14.

Tiwari, A. K. (2015). A Review on Big Data Integration. *International Journal of Computer Applications*. 1(1), 1–5.

Tomar, D. and Agarwal, S. (2014). A survey on pre-processing and post-processing techniques in data mining. *International Journal of Database Theory & Application*. 7(4), 99–128.

Torres, E. N., Singh, D. and Robertson-Ring, A. (2015). Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry. *International Journal of Hospitality Management*. 50, 77–83.

Tripadvisor (2016). 2015 Annual Report and First 2016 Meeting Notice.

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2013). Data mining with big data. *IEEE transactions on knowledge and data engineering*. 26(1), 97–107.

Xing, E. P., Jordan, M. I., Karp, R. M. *et al.* (2001). Feature selection for high-dimensional genomic microarray data. In *ICML*, vol. 1. Citeseer, 601–608.

Xu, X. and Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International journal of hospitality management*. 55, 57–69.

Yu, L. and Liu, H. (2003a). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*. 856–863.

Yu, L. and Liu, H. (2003b). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *International Conference on Machine Learning (ICML)*. ISBN 1577351894. ISSN 01469592, 1–8. doi:citeulike-article-id: 3398512. Retrievable at http://www.aaai.org/Papers/ICML/2003/ICML03-111.pdf .