A Data Analytics Approach to Online Tourists' Reviews Evaluation

Evripides Christodoulou[1], Andreas Gregoriades[1] and Savvas Papapanayides[2]
[1]*Cyprus University of Technology, Limassol, Cyprus*
[2]*American University of Bahrain, Bahrain*

Keywords: Sentiment Analysis, Ordinal Logistic Regression, Topic Analysis, Tourists' Reviews.

Abstract: This paper utilizes online data of tourists' reviews from TripAdvisor to identify patterns with regards to sentiment and topics discussed by tourists that visit Cyprus, along with the investigation of the effect of tourist culture and purchasing power on reviews' polarity, using logistic regression. The analysis uses natural language processing using the LDA technique and Naïve Bayes sentiment analysis. For the data collection, custom-made python scripts were used. Ordinal logistic regression is used to identify differences among the types of tourists visiting Cyprus, in accordance to culture and purchasing power.

1 INTRODUCTION

With the recent information explosion from the proliferation of data from the web, mobile apps, social media, and sensor networks, a new challenge emerged for companies to discover information patterns hidden in big data using effective data mining (Khade, 2016). A significant amount of data on the web relates to consumer evaluations. This active role of consumers in evaluating products and businesses through social media is changing organizations reputation (Etter, Ravasi and Colleoni, 2019) and sales (Rosario *et al.*, 2016) and has many practical applications in the area of marketing. The diffusion of consumers opinions in social media is often linked with emotions,(Berger and Milkman, 2012)(Pfeffer, Zorbach and Carley, 2014), that can affect company reputation and performance. Therefore, social media analytics is becoming a mainstream activity in marketing and is considered as a valuable tool in the evaluation and prediction consumers' behavior. Micro blogs are small messages communicated via social media such as Twitter, and gained popularity as means of expressing peoples' views (Chamlertwat *et al.*, 2012). Micro-blogs are also referred as an electronic word of mouth (eWOM), and constitute one type of big data with unstructured information. Companies analyze eWOM as part of their marketing strategy (Jansen *et al.*, 2009) to better position their products based on customer needs and opinions (Jung, 2008). According to Nayab, Bilal and Shrafat (Nayab, Bilal and Shrafat, 2016) a brand is no longer what the company tells a customer it is - it is, rather, what customers tell each other it is. Therefore, eWOM plays an important role in evaluating customers' perception of a brand or product. TripAdvisor and other social media platforms became valuable sources for eWOM analytics with techniques for mining consumers' sentiment and opinions. Several studies investigated the use of social networks to mine consumer-sentiment for customer behavior analysis (Moon and Kamakura, 2017) and product or business positioning (Lee, Rim and Lee, 2019) given evidence that sentiment in eWOM is a strong predictor of product success (Nguyen and Chaudhuri, 2019). However, they fail to analyze sentiment in the context of other parameters that have been identified as critical to consumers' emotion such as GDP and culture.

Therefore, this paper investigates the effect of culture and purchasing power of tourists on reviews polarity. The evaluation of reviews' sentiment is achieved using a Naïve Bayes sentiment classifier. The topics that each review discussed are identified using the LDA topic modelling approach. The main research questions addressed in this paper are:

1. How does purchasing power affects reviews' sentiment?
2. How does culture influence reviews' sentiment?
3. How reviews' discussion topics are linked with sentiment?

Christodoulou, E., Gregoriades, A. and Papapanayides, S.
A Data Analytics Approach to Online Tourists' Reviews Evaluation.
DOI: 10.5220/0009361000990105
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 1*, pages 99-105
ISBN: 978-989-758-423-7

The first question is grounded on evidence that purchasing power affects reviews' polarity, with consumers from countries with lower purchasing power providing low ratings to hotels. The second question is based on evidence that tourist cultural values, such as power distance, individualism, and uncertainty avoidance, significantly affect their perception of service quality, service evaluation, and satisfaction(Kim and Aggarwal, 2016).Their work however used the scenarios approach and hence limit the general is ability of their findings. Similarly other studies used surveys, to examine how customer power distance affects service expectations, perceived service quality, and relationship quality (Dash, Bruning and Acharya, 2009). Surveys however might be biased due to the sample used. Other similar studies highlight that in countries with greater power distance, customers feel superior to service providers (Kim and Aggarwal, 2016) and expect high service quality. This is linked to evidence that purchasing power (Schaninger, 1981)is linked with a greater need to portray status through consumption(Dubois and Duquesne, 1993), hence promoting power distance. The third research question is grounded on the importance of topics in reviews for the classification of issues that need attention (Nikolenko, Koltcov and Koltsova, 2017). All these influences however are analysed independently from each other; hence, this paper combines topic modelling with GDP and culture using regression to evaluate eWOM sentiment. This overcomes the problems of surveys that can suffer from limited sample size and sample bias.

The paper is organised as follows. The next section addresses the literature of sentiment and topic analysis along with literature pertaining to culture and purchasing power. Next section elaborates on the method followed and the results obtained. The paper concludes with the implication of the research and future directions.

2 LITERATURE REVIEW

2.1 Sentiment Analysis

Sentiment analysis (SA) and opinion mining have been studied for more than two decades with several techniques emerging during this time for analysing emotions and opinions from eWOM(Martin-Domingo, Martín and Mandsberg, 2019). SA is useful for online opinions analysis due to its ability to automatically measure emotion in online content using algorithms to detect polarity in eWOM(Pang and Lee, 2008).Three common SA approaches are: Machine Learning (ML), Lexicon-based Methods and Linguistic Analysis techniques. From the above three categories, the ML techniques are considered the most effective and simplest to use, with Naïve Bayes and Support Vector machines being the most popular. ML techniques can be either supervised or unsupervised (Witten *et al.*, 2016). As these are supervised learning techniques, it is important to train the classifiers prior to their use. The main difference from unsupervised is that supervised techniques use labelled opinions that have been pre-evaluated as negative, positive or neutral to train models. Such techniques include, Support Vector machines, Naïve Bayes, Logistic regression, Multilayer perceptron, K-Nearest Neighbours and Decision Trees(Krouska, Troussas and Virvou, 2017).

2.2 Topic Modelling

Topic modelling constitutes a popular tool for extracting important themes from unstructured data. It falls under the category of unsupervised data mining techniques employed to reveal and annotate large documents with thematic information (Nikolenko, Koltcov and Koltsova, 2017). Two of the most popular techniques for topic analysis are the Latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA)(Gambhir and Gupta, 2017). In LDA, a topic is a probability distribution function over a set of word, used as a type of text summarization. The PLSA approach expresses the relationships between words in terms of their affinity to certain hidden variables (topics), just as in LDA, unlike LDA though, this relationship is expressed in probabilities, instead of Dirichlet prior probabilities. LDA, is a Bayesian version of PLSA and has better generalization. Therefore, LDA is employed in this study, with each review representing a distribution of a finite set of topics, each one being a multinomial distribution of the words in the corpus that is developed from all reviews. LDA examines a collection of reviews and learns what words tend to be used in similar reviews to identify the main topics in the corpus.

2.3 Culture

A key factor that differentiates tourist activities is their culture, with studies such as Crotts and Erdmann (2000) identifying that certain traits have significant effect on tourist satisfaction during a visit to a country. People of the same nationality tend to have analogous preferences and similarities in their consumer behavior(Huang and Crotts, 2019). There are several models of culture. In this study we adopted the model of Hofstede(2011)due to its reputation. According to this model there are 6 different traits that form a culture: Power Distance - The degree to which people accept and expect that power is distributed unequally in a country. Individualism- when people tend to take care of only themselves and their immediate families. Masculinity - where achievement, heroism, assertiveness, and material rewards for success are preferred in a society. Uncertainty Avoidance - when risk and uncertainty tend to be avoided. Long Term Orientation - when people prefer stability, respect for tradition, and are future-oriented. Indulgence -when people prefer freedom and free will.

For the purpose of this study, we have conducted in-depth analysis based on Hofstede cross cultural differences model, focusing on how specific traits affect sentiment in online tourist reviews.

2.4 Purchasing Power

Another important variable that varies from country to country and is not included in the elements of culture, is the financial state of the tourist country of origin in relation to that of Cyprus. Purchasing power has been used extensively for global market analysis (Gilboa and Mitchell, 2020). The economic performance of a country does not only represent its financial status but is also related to people's purchasing behavior either within the country or outside. Gross Domestic Product (GDP) per capita is one key indicator for comparing the level of development among countries and is also used as a socioeconomic indicator of health. It is widely considered that human welfare and GDP per capita go together, while increased GDP per capita is correlated with happiness of people(Dipietro and Anoruo, 2006). At the same time in countries with low human development index, GDP dramatically changes quality of life(Islam, 1995). Therefore, the hypothesis here is that tourists visiting Cyprus from countries with lower purchasing power compared to Cyprus, are most likely to be more demanding and hence more likely to evaluate the hotel and its services negatively.

3 METHODOLOGY

The main steps required to answer our research question are the following. The first step is the collection of tourist reviews from all hotels in Cyprus for the period 2009-2019. The total number of reviews was 65000 from tourists coming from 27 countries, stayed at 2 to 5 stars hotels and the language of review was English. In this step, an automated technique is used to collect the data based on specific criteria.

The data collected underwent pre-processing, that involved data cleansing, dimensionality reduction (clustering of GDP values was performed due to scarcity of data among 27 countries) and irrelevant data elimination. Pre-processing is a necessary step that improves data quality. The next step involves the analysis of consumers' sentiment and the topics of eWOM through polarity detection and topic analysis. Sentiment analysis is required to detect cases when reviewers' rating is neutral, but the actual text contains negative connotations. For the sentiment analysis, a Python algorithm was developed to train a Naïve Bayes classifier using the "nltk" library, to evaluate the polarity of reviews in three categories: positive, negative and neutral. For the topic identification the LDA approach is utilized due to its popularity and proven results. The final step in the method addresses the longitudinal effect of culture and purchasing power to reviews sentiment. This is evaluated using ordinal logistic regression.

LDA pre-processing step refers to the procedure of cleansing and preparing reviews that are going to be analysed. Unstructured information on the Internet contains significant amounts of noise, such as data that do not contain any useful information for the analysis at hand. Filtering irrelevant information preceded the analysis, to eliminate useless metadata (ascii characters or URLs). The pre-processing involved the steps of: cleansing stop-word removal, tokenisation, stemming, lemmatisation and filtering. Stop-word refers to words providing little or no useful information to text analysis and can hence be considered as noise. Common stop-words include articles, conjunctions, prepositions, pronouns, etc. Other stop-words are those typically appearing very often in sentences, or in specific contexts. Tokenization refers to the transformation of a stream of strings into a stream of processing units, referred to as tokens. Thus, during this step reviews were converted into a sequence of tokens, by choosing n-grams (phrases composed by n words in length) as tokens after removing punctuation marks and special symbols. Stemming and lemmatization processes involved converting a word to its root form and is typically required in dealing with fusional languages, like English. Lemmatization uses a vocabulary and morphological analysis of words, to return the base-form of a word, known as the lemma. Lemmatization, unlike stemming, reduces the word to its lemma, ensuring that the root word belongs to the language and context of interest. Stemming usually employs a heuristic process that eliminates endings of words which often results in the removal of derivational affixes. This process is sometimes called word normalization in NLP, and consists of reducing each token to its stem, in order to group words having closely related semantics. For instance, “Playing”, “Plays” and “Played” become “Play”. Filtering involved the removal of words(stems) considered as irrelevant such as names of individuals. Thus, each review is cleaned from stems not belonging to the set of relevant stems.

The LDA model is learned using the Gibbs sampling technique that essentially performs a random walk in a way that reflects the characteristics of the desired distribution, starting at a random initial point. To improve the comprehension of the generated model, the terms in each topic are ranked based on their frequency. This is expressed by the beta values that are the Dirichlet priors for tokens over topics. Extracted topics were inspected based on prior domain knowledge, therefore, expertise in the field under investigation is required to make the necessary connections. The refined number of topics for the final LDA model was 8, after evaluating results of various topic sizes based on the estimated number of k. Following the learning of the LDA model and the identification of the main topics, each review was associated with a topic(s) based on the trained LDA model and the result was saved in a new datafile. To further refine the topics that emerged from the analysis, an ontology was utilised, defined based on domain knowledge (Dickinger and Mazanec, 2008)that represented the main dimensions that reviewers in the tourism industry refer to. Therefore, this ontology was used to group certain topics together to form sub-topics that better related to the hospitality industry. The sub-topics that emerged were: Location (area, shops, nearby), Facilities (pool, water sport, bar) , Service (superb, professional, staff), Money (price, cost, room), Accessibility (disable, lift, room). To make this association, a Python script was used that associated topics that referred to words linked with each of the dimensions and assigned one or more sub-topics to each review in the dataset.

To analyse the effect of culture and purchasing power on sentiment, it was imperative to augment the dataset with relevant cultural and purchasing power information based on the country of origin of the author of each review. For the estimation of the purchasing power of each reviewer, the ratio of GDP between the visitor’s country and Cyprus was utilised. The data for the GDP of each country was obtained from the world momentary fund website. Similarly, for the association of each reviewer’s country with the relevant cultural dimensions (section 2.2), the Hofstede website was used in order to express the cultural dimension of a country in a scale from 0-100.

3.1 Data Collection

To extract data from TripAdvisor, an algorithm was developed in Python that scrapped reviews of tourists that visited Cyprus in the period 2009-2019. The data collected included: Username, Rating of hotel, Date of stay, Feedback date, Country of origin, Pas Contributions, Confidence votes, Review. This work focused on review text, county of origin and date. An initial visualisation of the data is depicted in Figure 1, showing the distribution of sentiment polarity with time.

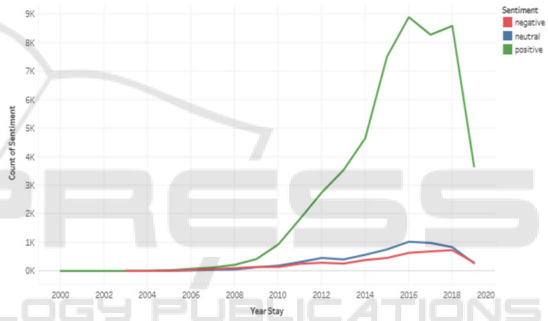

Figure 1: Distribution of sentiment polarity and time.

4 RESULTS

The Naïve Bayes classifier was trained using the reviewers rating of the hotel as an indication of their sentiment. So high rating was associated with positive sentiments and low rating with negative sentiments. The performance of the trained model was compared against two pre-trained models: Textblob and Vader (Hutto and Gilbert, 2014), which are popular alternatives with satisfactory precision and recall scores. Textblob and Vader are based on bag of words method, but the former also includes subjectivity analysis estimates. The metric of subjectivity is in the range of [0-1] with 1 referring to subjective and 0 to objective content. Both classifiers performed similarly to the trained model, hence were used in an ensemble manner to improve our confidence in the results. The developed Python algorithm automatically utilized the Textblob and Vader models along with the trained Naïve Bayes model and averaged their results. The process was repeated for all downloaded reviews, and their polarity and subjectivity were saved next to the review in a new csv datafile.

The 65000 reviews then underwent an initial descriptive analysis revealing approximately the following distribution of review sentiment by polarity: 10% negative 10% neutral and 80% positive. Additional descriptive statistics (Figure 2) revealed that Paphos is the town with the most reviews and the town with the most neutral and negative sentiments in its review, while Famagusta is the one with the most positives reviews.

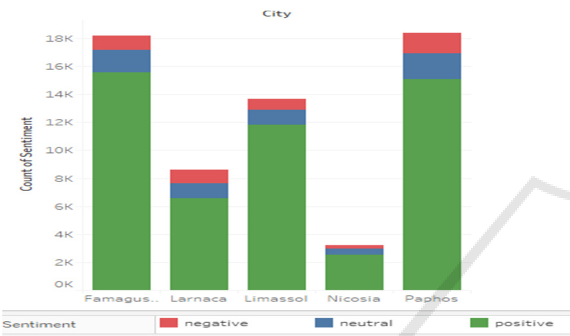

Figure 2: Distribution of review frequency per town and sentiment polarity.

4.1 Empirical Model

To examine the effect of the independent variables (culture, purchasing power) on tourist sentiment, an ordinal logistic regression (OLR) model was specified with sentiment being the dependent variable and culture/purchase power respectively the independent variables. The OLR model aimed to identify how well the independent variables predict the ordinal dependent variable. The SPSS statistical tool was used to estimate the effect of each cultural dimension and purchasing power on reviews' sentiment. Ordinal logistic regression takes ordinal variables as dependent variables and scale or category variables as independent. This technique enables the estimation of the probability of the independent variable affecting the dependent variable. There are several OLR models such as proportional odds model, two versions of the partial proportional odds model-without restrictions and with restrictions, continuous ratio model, and stereotype model. The most popular model is the proportional odds model used here.

To estimate each country's purchasing power the gross domestic product (GDP) was used, based on the World Monetary Fund dataset. The original dataset was expressed in US dollars; hence, these were converted to Euros to enable the comparison with the GDP of Cyprus. The results of dividing the GDP of each country with the GDP of Cyprus, enables comparison of the purchasing power of each tourist's country of origin to that of Cyprus.

The OLR analysis performed in this study used categorical data for purchasing power (GDP) to group countries of origin into clusters. The transformation of the input numerical values of purchasing power into new categories was performed based on characteristics of the 6 main clusters that emerged after conducting k-means clustering on all countries purchasing power.

Therefore, the original dataset was recoded based on these new 6 categories, based on their purchasing power. The first category with code 1 refers to GDP ratio to Cyprus under 0.6, category 2 [1.5 to 2.4], category 3 [2.5 to 3.4],category 4, [3.5 to 4.4], category 5 from 4.4+ and category 6 from [0.6 to 1.4]has been used as a reference category.

Table 1: Effect of GDP Ratio on Sentiment from OLR. Richer countries more likely to give positive reviews compared to poorer countries.

GDR Ratio	Estimate	Significance
[GDP_A=1]	0.221	0.00
[GDP_A=2]	0.453	0.00
[GDP_A=3]	0.261	0.012
[GDP_A=4]	0.101	0.628
[GDP_A=5]	-0.367	0.77
[GDP_A=6]	Reference category (Cyprus)	

To answer the first research questions, the OLR was used to find the relationship between tourists purchasing power, on sentiment. Table 1 shows that the model's coefficient of certain countries purchasing power are significant ($p < 0.05$), thereby suggesting that the reviewers' country of origin is related to their online hotel ratings. The reviewers with higher purchasing power tend to leave positive reviews.

To investigate the effect of cultural traits on tourists' review sentiment, the Hofstede insights website was used to assign each country's cultural dimension to all reviews. Culture metrics are divided into 6 categories on a scale ranging from 0 to 100. The traits as mentioned before are power distance, individualism, motivation for success and masculinity, uncertainty avoidance, long term orientation and lastly indulgence. Results from the effect of culture on sentiment are depicted in Table 2.

Table 2: Effect of Culture on Sentiment from OLR. Power distance and uncertainty avoidance having a negative effect on sentiment, while individualism having a positive effect.

Cultural trade	Estimate	Significance
Powerdistance	-0.002	0.069
Uncertaintyavoidance	-0.004	0.001
Individualism	0.004	0.015
Masculinity	0.001	0.522
Longtermorientation	0	0.848
Indulgence	-0.007	0

Finally, to investigate which topics had significant effect on sentiment, we utilised the results of the LDA model from previous step and combined subtopics in all possible permutations. The combinations of sub-topics that yielded significant results is depicted in Table 3.

Table 3: Effect of Sub-Topic combinations on Sentiment from OLR.

Sub-Topics Combinations	Estimate	Significance
[5]	0.718	.024
[1,2,5]	0.797	.013
[1,2,3]	1.566	.000
[1,2,3,5]	1.536	.000
[1,2,3,4]	0.888	.006
[1,2,3,4,5]	1.264	.001
[2,3]	1.566	.000
[2,3,5]	1.536	.000
[1]	0.827	.014
[1,3]	1.566	.000
[2,3,5]	1.536	.000
[3]	0.888	.000
[3,5]	1.264	.002

The used subtopics refer to: locations (1), facilities (2), service (3), money (4), accessibility (5). Combinations of these were used as the predictors of sentiment in the OLR model. Results highlighted that, the topics that are the most influential to positive sentiment, by the reviewers, were the ones that included the following combinations of subtopics: location of the hotel, the level of service and the accessibility of the venue. Therefore, if the hotel is at a good location, is easily accessible and provides good service, the likelihood that it will be evaluated positively is increased.

5 CONCLUSIONS

This study investigated the influences of culture dimensions and purchasing power on online hotel reviews, from TripAdvisor. Four critical findings are obtained. First, consumers from countries with lower purchasing power provide low ratings to hotels; this finding is consistent with similar studies that evaluate the power distance difference of tourist from different countries and how it affects online reviews. This is based on theory highlighting that in countries with high power distance, inequalities are generally accepted by individuals (Hofstede, 2011)and consumers often feel superior to service providers in the social hierarchy (Kim and Aggarwal, 2016),and expect high service quality while they tend to give low service evaluations.

Results from this work extends above findings with evidence that other cultural traits from Hofstede, such as individualism and uncertainty avoidance, tend to affect tourist review sentiment, while the topics that are associated with highest sentiment are those associated with service, location and accessibility of the hotel, indicating that the facilities of hotels in Cyprus are perceived by tourists as satisfactory and hence are evaluated with positive sentiment.

Limitations of this work resides in the quality of the data collected and issues pertaining fake reviews that might affect the results. Our future work aims to filter out these reviews and examine if the effect of the variables is altered.

REFERENCES

Berger, J. and Milkman, K. L. (2012) 'What Makes Online Content Viral?', *Journal of Marketing Research*, 49(2), pp. 192–205.

Chamlertwat, W. *et al.* (2012) 'Discovering Consumer Insight from Twitter via Sentiment Analysis', *Journal of Universal Computer Science*, 18(8), pp. 973–992.

Dash, S., Bruning, E. and Acharya, M. (2009) 'The effect of power distance and individualism on service quality expectations in banking: A two-country individual- and national-cultural comparison', *International Journal of Bank Marketing*, 27(5), pp. 336-358 ·.

Dickinger, A. and Mazanec, J. (2008) 'Consumers' Preferred Criteria for Hotel Online Booking', in *Information and Communication Technologies in Tourism 2008*, pp. 244–254.

Dipietro, W. R. and Anoruo, E. (2006) 'GDP per capita and its challengers as measures of happiness', *International Journal of Social Economics*, 33(10), pp. 698–709.

Dubois, B. and Duquesne, P. (1993) 'The Market for Luxury Goods: Income versus Culture', *European Journal of Marketing*, 27(1), pp. 35–44.

Etter, M., Ravasi, D. and Colleoni, E. (2019) 'Social Media and the Formation of Organizational Reputation', *Academy of Management Review*, 44(1), pp. 28–52.

Gambhir, M. and Gupta, V. (2017) 'Recent automatic text summarization techniques: a survey', *Artificial Intelligence Review*, 47(1), pp. 1–66.

Gilboa, S. and Mitchell, V. (2020) 'The role of culture and purchasing power parity in shaping mall-shoppers' profiles', *Journal of Retailing and Consumer Services*, 52.

Hofstede, G. (2011) 'Dimensionalizing Cultures: The Hofstede Model in Context', *Online Readings in Psychology and Culture*, 2(1).

Huang, S. (Sam) and Crotts, J. (2019) 'Relationships between Hofstede's cultural dimensions and tourist satisfaction: A cross-country cross-sample examination', *Tourism Management*, 52, pp. 232–241.

Hutto, C. J. and Gilbert, E. (2014) 'VADER: A parsimonious rule-based model for sentiment analysis of social media text', in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Islam, S. (1995) 'The human development index and per capita GDP', *Applied Economics Letters*, 2(5), pp. 166–167.

Jansen, B. J. *et al.* (2009) 'Twitter power: Tweets as electronic word of mouth', *Journal of the American Society for Information Science and Technology*, 60(11), pp. 2169–2188.

Jung, J. J. (2008) 'Taxonomy alignment for interoperability between heterogeneous virtual organizations', *Expert Systems with Applications*, 34(4), pp. 2721–2731.

Khade, A. A. (2016) 'Performing Customer Behavior Analysis using Big Data Analytics', *Procedia Computer Science*, 79, pp. 986–992.

Kim, C. S. and Aggarwal, P. (2016) 'The customer is king: culture-based unintended consequences of modern marketing', *Journal of Consumer Marketing*, 33(3), pp. 193–201.

Krouska, A., Troussas, C. and Virvou, M. (2017) 'Comparative evaluation of algorithms for sentiment analysis over social networking services', *Journal of Universal Computer Science*, 23, pp. 755–768.

Lee, H.-C., Rim, H.-C. and Lee, D.-G. (2019) 'Learning to rank products based on online product reviews using a hierarchical deep neural network', *Electronic Commerce Research and Applications*, 36, p. 100874.

Martin-Domingo, L., Martín, J. C. and Mandsberg, G. (2019) 'Social media as a resource for sentiment analysis of Airport Service Quality', *Journal of Air Transport Management*, 78, pp.106-115.

Moon, S. and Kamakura, W. A. (2017) 'A picture is worth a thousand words: Translating product reviews into a product positioning map', *International Journal of Research in Marketing*, 34(1), pp. 265–285.

Nayab, G., Bilal, M. and Shrafat, A. S. (2016) 'A brand is no longer what we tell the customer it is - it is what customers tell each other it is: Validation from Lahore, Pakistan', *Science International (Lahore)*, 28(3), pp. 2725–2729.

Nguyen, H. T. and Chaudhuri, M. (2019) 'Making new products go viral and succeed', *International Journal of Research in Marketing*, 36(1), pp. 39–62.

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

Pang, B. and Lee, L. (2008) 'Opinion mining and sentiment analysis', *Foundations and Trends in Information Retrieval*, 2(1–2), pp. 1–135.

Pfeffer, J., Zorbach, T. and Carley, K. M. (2014) 'Understanding online firestorms: Negative word-of-mouth dynamics in social media networks', *Journal of Marketing Communications*, 20(1–2),pp. 117–128.

Rosario, A. B. *et al.* (2016) 'The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors', *Journal of Marketing Research*, 53(3), pp. 297–318.

Schaninger, C. M. (1981) 'Social Class versus Income Revisited: An Empirical Investigation', *Journal of Marketing Research*, 18(2), pp. 192–208.

Witten, I. H. *et al.* (2016) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann