UCL DEPARTMENT OF GEOGRAPHY

YEAR 2022-23

EXAM CANDIDATE ID:	**BCHJ0**
MODULE CODE:	**GEOG051**
MODULE NAME:	**Mining Social and Geographic Datasets**
COURSE PAPER TITLE:	**Mobility Analysis Using Gowalla Data and Shopping Mall Location Selection / Predicting Customer Sentiments in Calgary Using Unstructured Review Data**
WORD COUNT:	**2492**
CODE REPOSITORY LINK:	**https://anonymous.4open.science/r/data-mining-coursework-BCHJ0**

Your essay, appropriately anonymised, may be used to help future students prepare for assessment. Double click this box to opt out of this ☐

Mobility Analysis Using Gowalla Data and Shopping Mall Location Selection

Introduction

With smartphones capable of retrieving geographical information, location-based services have become prevalent (Rallapalli *et al.*, 2013). With the initial use of Facebook in 2007, Location-Based Social Media Networks (LBSNs) have become part of our lives for commercial, governmental, and non-profit purposes (Ullah *et al.*, 2022). LBSNs, such as Foursquare and Gowalla, were so popular at the beginning of the 2010s, enabling a wide range of locational data 'voluntarily' generated by the users despite serious privacy concerns.

Check-in data generated in Foursquare and Gowalla can be used in mobility analysis to understand the movement of people, such as finding the shortest route between check-in locations and even predicting people's next location based on previous ones. Pellet, Shiaeles and Stavrou (2019) discussed privacy issues related to LBSN data by showing how well a user's location can be predicted by machine learning.

Mobility is also essential for the location selection of land uses while planning a city, together with other traditional theories, such as central place theory and land value theory (Dawson, 2013, cited in Lin, Chen and Liang, 2018). Street centrality is vital in selecting the locations of retail stores, particularly for shopping malls, due to its impact on land use and socioeconomic activities (Lin, Chen and Liang, 2018).

The objective of this study is twofold: (1) to find the shortest driving paths between each consecutive check-in point from Gowalla for two specific users in Cambridge, UK, while exploring the spatial distribution and referring to privacy concerns, and (2) to measure closeness centrality and propose the location of a shopping mall in Cambridge.

Data and Methods

The data from Gowalla contains individual check-in details, such as user ID, latitude, longitude, date and time. Firstly, the tabular data obtained is converted into geospatial data using latitude and longitude. Check-in locations of two users, one on 30th January and the other on 24th May 2010 in Cambridge are selected to find the shortest driving routes per user using OpenStreetMap network data.

The street centrality is calculated based on the closeness centrality measure with OpenStreetMap data to propose a new shopping mall in an accessible place. Closeness centrality, which represents the advantage of overcoming spatial separations between locations, is measured by identifying the closeness of each node to all nodes along the shortest pathways (Yu, 2017). On Google Maps, there are approximately five shopping malls and comparably smaller retail stores in Cambridge. The locations of existing shopping malls are also considered while proposing a new one.

Results

Figure 1 shows that the mobility of one user was within the western side of Cambridge, while the other user moved to the eastern and southern parts of the city. This kind of spatiotemporal data can be used viciously against users due to the ease of being identified by multiple locations. Moreover, the daily route of a user can be calculated, and the next location can be predicted at a specific time of the day, and it can pose a physical threat as well.

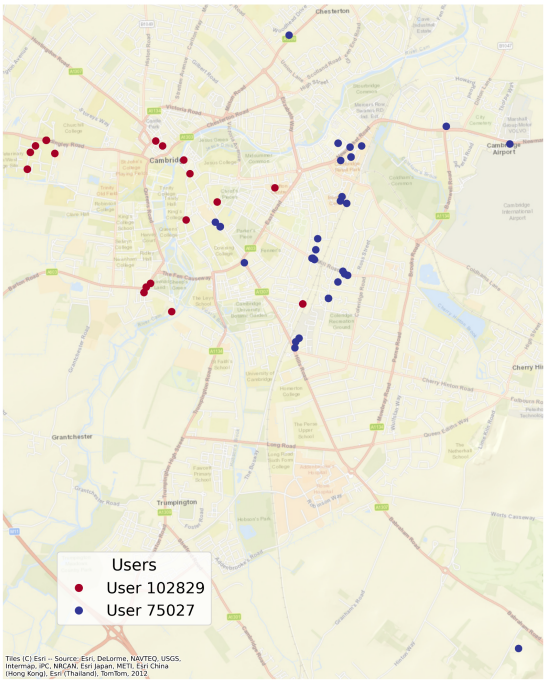

Figure 1. Individual Check-in Locations for Two Users

The calculated shortest driving path for each user is demonstrated in figure 2. Both users travel almost the same amount of road. However, the maximum displacement of user 75027 is higher, while the average displacement is lower than user 102829.

Figure 2. The Shortest Path for Users 75027 (left) and 102829 (right)

Users	Network Type	Maximum Displacement	Average Displacement	Total Travel
75027	Driving	7078 meters	1999 meters	15994 meters
102829	Driving	5147 meters	2247 meters	15730 meters

Table 1. Statistics of shortest routes per user

Figure 3 shows a new shopping mall location in the northern part of the city, where there is an essential road with higher closeness connectivity. The existing shopping malls are in the central part of the city. The urban development pattern spreads through the southeastern part of the city. This part of the city has lower closeness connectivity and already has urban development pressure. The proposed location does not have an agricultural area, as seen from the imagery basemap.

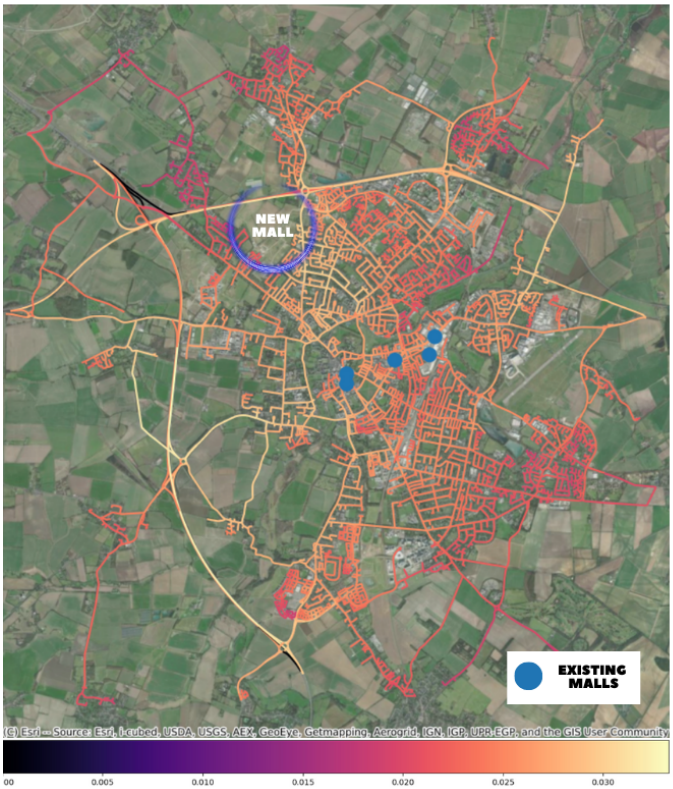

Figure 3. Closeness Connectivity in Cambridge

Conclusion

The study shows that data from Gowalla can be used to find the shortest routes. However, it can pose genuine threats for users, ranging from identification to location prediction. Although the popularity of social media platforms, such as Gowalla, has decreased, the other popular ones should also take privacy and data protection issues into consideration, particularly for locational data.

The study also proposes the location for a new mall based on closeness centrality, a little bit of consideration of land use and existing malls, together with the urban development pattern of Cambridge. Yet, a new shopping mall can create problems for local businesses and create development pressure as well.

Bibliography

Lin, G., Chen, X. and Liang, Y. (2018) 'The location of retail stores and street centrality in Guangzhou, China', *Applied Geography*, 100, pp. 12–20. Available at: https://doi.org/10.1016/j.apgeog.2018.08.007.

Pellet, H., Shiaeles, S. and Stavrou, S. (2019) 'Localising social network users and profiling their movement', *Computers & Security*, 81, pp. 49–57. Available at: https://doi.org/10.1016/j.cose.2018.10.009.

Rallapalli, S. *et al.* (2013) 'Analysis and applications of smartphone user mobility', in *2013 Proceedings IEEE INFOCOM*. *2013 Proceedings IEEE INFOCOM*, pp. 3465–3470. Available at: https://doi.org/10.1109/INFCOM.2013.6567182.

Ullah, H. *et al.* (2022) 'Understanding the User-Generated Geographic Information by Utilizing Big Data Analytics for Health Care', *Computational Intelligence and Neuroscience*, 2022, p. e2532580. Available at: https://doi.org/10.1155/2022/2532580.

Yu, W. (2017) 'Assessing the implications of the recent community opening policy on the street centrality in China: A GIS-based method and case study', *Applied Geography*, 89, pp. 61–76. Available at: https://doi.org/10.1016/j.apgeog.2017.10.008.

Predicting Customer Sentiments in Calgary Using Unstructured Review Data

Introduction

The availability of data regarding products/services has increased thanks to user-generated content on the internet. Online reviews on e-commerce websites and other social media platforms are undoubtedly helpful for customers wishing to find suitable products (Bansal and Srivastava, 2018) and companies trying to understand customer preferences. Customers care about other customers' opinions while purchasing (Poushneh and Rajabi, 2022). 93% of customers make decisions based on online reviews, according to a recent study (Kaemingk, 2019, cited in Schoenmueller, Netzer and Stahl, 2020). Customer satisfaction rating and review writing are the most prevalent ways customers express their sentiments. Although analysing rates is comparably more manageable, text reviews are unstructured and need advanced mining methods to extract insights for further analysis (Jia, 2018).

With advances in natural language processing (NLP), deriving meaningful information from unstructured textual data, such as quantifying customers' opinions on products or services, has become possible. NLP is a technique that enables the communication between computer programmes and human language. NLP provides many text-mining methods for understanding the relationship between online reviews and ratings (Jia, 2018). Predicting customer satisfaction ratings using textual data has been particularly vital for companies since reading thousands of data requires time and human resources. Zahoor, Bawany and Hamid (2020) used online reviews of customers about restaurants in Karachi, Pakistan, to predict the polarity of sentiments and category classification. They found that Random Forest gives the best accuracy for both sentiment and category classification, among other machine learning methods.

The objective of this study is twofold: (1) to understand if extracted information from unstructured review data using text mining can explain the polarity (positive or negative) of sentiments derived from customer satisfaction rates in Calgary, Canada, and (2) to explore the spatial distribution of polarity across Calgary, particularly for restaurants since 55% of all reviews are generated for restaurants in the dataset. The location is proposed for a new potential restaurant based on polarity distribution as well.

Data and Methods

This study uses online review data from a social media platform, with 15 columns including customer rates, reviews, categories and location information. There are 82,182 reviews that 23,925 different users generated between 2008 and 2019. This study focuses on ten years between 2010 and 2019 since old reviews may not reflect the current situation. Moreover, the average number of reviews for each venue in the dataset is six. However, some businesses have an extremely high number of reviews. Therefore, 5797 outliers of review counts were removed from the dataset. Table 1 shows the entire data-wrangling process of the study. Ultimately, 74,621 reviews were used in the analysis. Each review receives a corresponding rating between 1 and 5. The data has been classified into positive or negative categories based

on customer rates. This polarity variable is the dependent variable modelled to be predicted using reviews.

Variables	Process	Results	Usage in the Anaysis
Business ID	Grouping businesses	3014 Businesses	Finding average rate for each venue and exploring distribution
Review Counts for Each Venue	Removing outliers	5797 rows removed	-
User ID	Grouping users	23,925 users	-
Latitude/ Longitude	Geolocating venues	Geodataframe	The spatial distribution of business and restaurants
Categories	Feature engineering	Four different columns	Filtering restaurants
Rating Star	Deriving polarity of customer satisfaction	A column indicating positive or negative sentiments	Derived polarity column as dependent variable
Review Text	Text mining	Bag-of-words excluding non-Latin characters, stop words, and punctuations	Bag-of-words as independent variables
Tags (Useful, cool, funny)	Concating with bag-of-words	A dataset ready for splitting train and test	Independent variables
Date	Deriving the year	1763 reviews in 2008 and 2009 removed	-

Table 1. Data Pre-Processing

The most tricky part of the study is to create a structured table with meaningful words from unstructured reviews. Therefore, non-Latin characters, stop words, extra whitespace, and punctuation were removed. Case conversion, lemmatisation and tokenisation are performed to derive bag-of-words (BOW). The BOW is a document term matrix representing text by conceptualising each term as a feature that appears in a document (Sánchez-Franco, Navarro-García and Rondán-Cataluña, 2019). The derived BOW, together with a number of tags (useful, funny, cool), are used as dependent variables to predict the polarity of customer satisfaction. Naive Bayes, Decision Tree and Random Forest Machine Learning methods are conducted to perform binary classification of positive or negative sentiments with training (80%) and test (20%) datasets.

To explore the spatial distribution of venues, the dataset is grouped by businesses. The polarity is identified for this grouped data based on average values. Regarding the type of venues, the restaurant is the most prevalent category, and 28% of all venues in the dataset are restaurants. The distribution of restaurants is particularly explored to propose a rough location for a new restaurant. Figure 1 shows the overall workflow of this study.

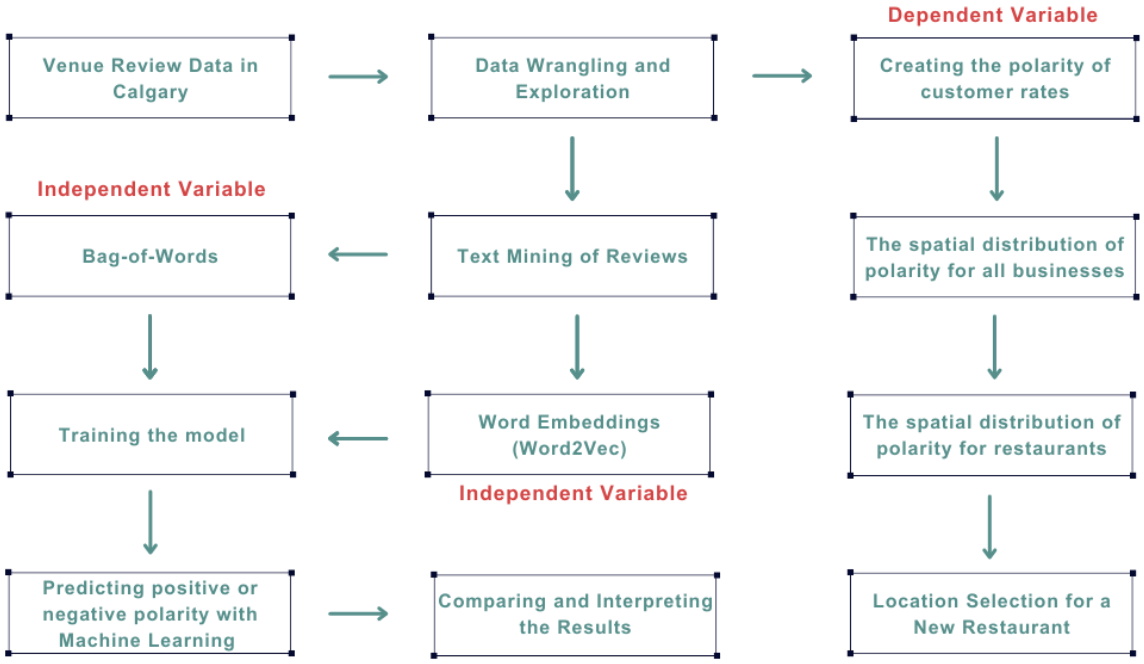

Figure 1. Research Pipeline

Results

The polarity of customer sentiments is derived by considering four and five-star rates as positive polarity and three-star and less as negative polarity. Figure 2 demonstrates that the large proportion of sentiments is positive. Customers often give five and four-star rates, followed by one, three and two-star consecutively. Textual reviews are used to predict this binary variable.

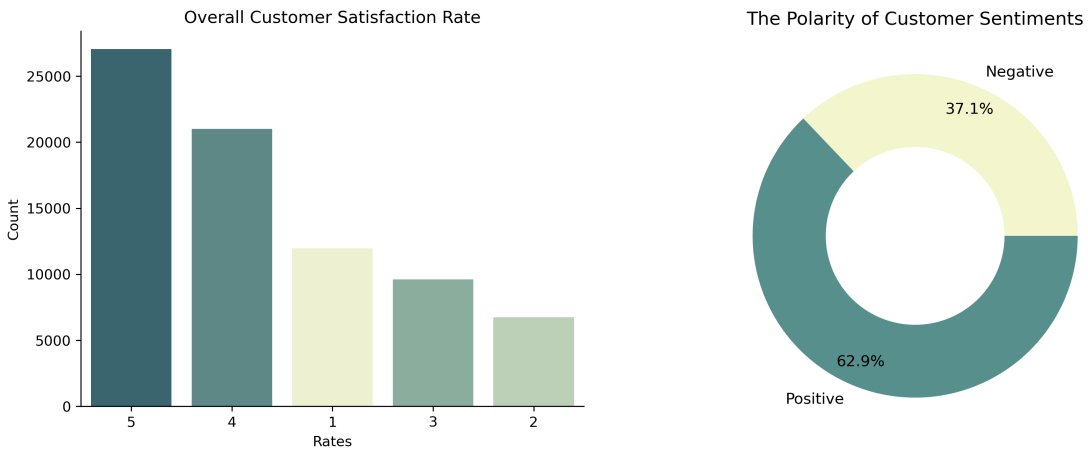

Figure 2. a) Customer Satisfaction Rate **b)** The Polarity

Besides structured textual data, the number of tags is also used as the independent variable to classify reviews as positive or negative. Three columns show the number of time the other user find the review useful, cool or funny.

Online reviews are cleaned and converted into BOW format for the model. Only 2500 most frequent words out of 32,295 are included in the BOW for computational efficiency reasons. However, the results remain the same as the model with all words. Figure 3 shows the frequent words in the BOW. Although more neutral words seem dominant, the words related to positive and negative sentiments can also be seen from the graph.

Figure 3. Word Cloud

BOW provides a format similar to one-hot encoding, with ones for existence and zeros for non-existence. However, this method cannot measure the similarity of words. Therefore, word embedding, namely Word2Vec, is also used. Word embedding transforms text into vectorised real numbers, and World2Vec, as a version of it, calculates similarities between words (Alharbi *et al.*, 2021). In this study, a pre-trained model of Google News 300 is used.

Decision Tree, Random Forest and Naive Bayes methods are conducted with three different sets of variables to classify the polarity of sentiment. Table 2 shows that all three perform well regarding the accuracy score. Random Forest algorithm with a 5-fold cross and grid search on hyperparameters has the best performance comparably with an accuracy of 84%. The results of Naive Bayes (82%) are almost identical as with Random Forest. Decision Tree yields the lowest accuracy score comparably with all three sets of variables. In fact, incorporating tag variables into the model does not increase the accuracy of models significantly. Random Forest exhibits the highest increase by 0.7%. Moreover, vectorising textual data with pre-trained word embeddings does not increase the accuracy of models as expected. There is a slight decrease in the accuracy of each model with word embedding.

Methods	Accuracy Score (Only BOW)	Accuracy Score (BOW + tags)	Accuracy Score (Word2Vec + tags)
Decision Tree	0.753	0.754	0.709
Random Forest	0.835	0.842	0.81
Naive Bayes	0.821	0.822	-

Table 2. Results of ML Classification

Although the accuracy scores of Random Forest and Naive Bayes were approximately the same, Table 3 shows that Random Forest predicts positive polarity (True Positive of 8964) more correctly than Naive Bayes. On the other hand, Naive Bayes is better at predicting negative polarity (True Negative of 3968). Decision Tree predicts slightly better negative polarity than Random Forest.

Methods	Decision Tree		Random Forest		Naive Bayes	
	Actual Positive	Actual Negative	AP	AN	AP	AN
Predicted Positive	7604	1779	8964	419	8305	1078
Predicted Negative	1879	3663	1939	3603	1574	3968

Table 3. Confusion Matrix

Figure 4 reveals a concentration of positive sentiments towards all venues and restaurants in the city centre of Calgary. However, negative sentiments towards all venues tend to be clustered

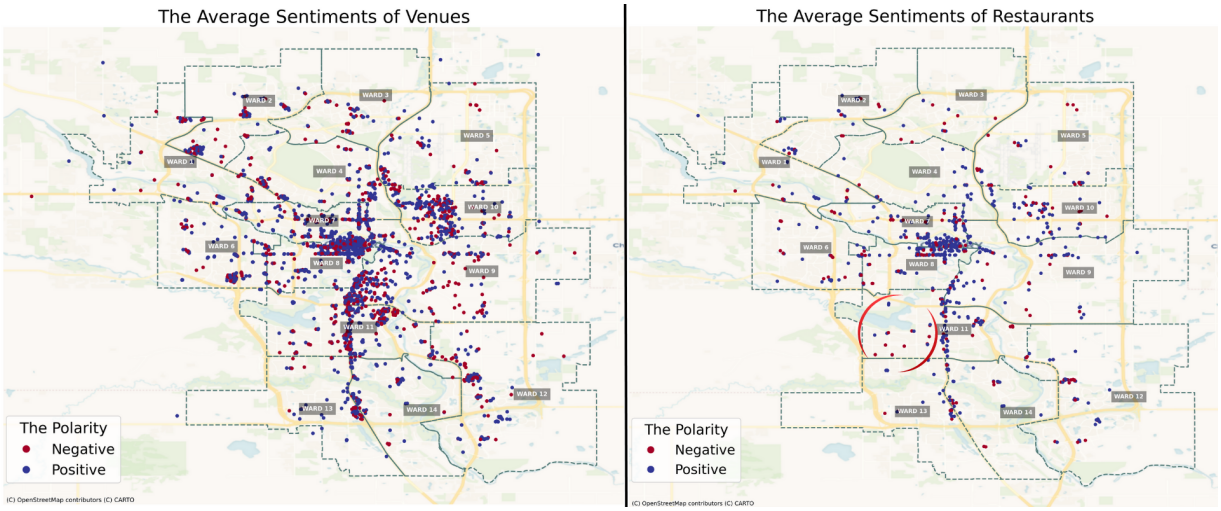

Figure 4. The spatial distribution of the polarity for **a)** all businesses **b)** restaurants

in Ward 9 and 11. For restaurants, the west of Ward 11 can be the concentration point of negative sentiments. Other wards do not have dominant polarity on the map. The western part

of Ward 11 also has an attraction point around Glenmore Reservoir with many recreational areas. Considering this attraction point and low sentiments toward existing restaurants, a new restaurant can be opened here.

Conclusion

The study shows that textual review data from social media can be useful in predicting the polarity of customer sentiments in Calgary. Overall, the results indicate all ML methods used in the analysis perform well in classifying binary sentiments. The number of tags each review received from other users does not contribute to the accuracy score of models as independent variables. Moreover, the model trained with BOW obtains a higher score than word embedding (Word2Vec) even though it considers the similarity and location of the word in the text.

The study also reveals a concentration of low sentiments around Glenmore reservoir and proposes a new location for a restaurant there, considering the competitive advantage one can have over unpopular restaurants and the potential need for a new restaurant due to recreational areas. The limitation of the study is related to the self-selection bias that occurs when users publish their reviews (Sánchez-Franco, Navarro-García and Rondán-Cataluña, 2019). Reviews may not have been written by real customers as well. Moreover, younger and those who have access to social media can be dominant among users. Therefore, the reviews may not be representative. Regarding the rates, users may have been affected by the average rate.

Business owners can easily evaluate sentiments without any detail reading the full reviews, although each review can provide genuine insight into how to improve customer satisfaction. Further studies can be conducted to see how sentiments have changed over time with temporal analysis.

Bibliography

Alharbi, N.M. *et al.* (2021) 'Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews', *Mathematical Problems in Engineering*, 2021, p. e5536560. Available at: https://doi.org/10.1155/2021/5536560.

Bansal, B. and Srivastava, S. (2018) 'Sentiment classification of online consumer reviews using word vector representations', *Procedia Computer Science*, 132, pp. 1147–1153. Available at: https://doi.org/10.1016/j.procs.2018.05.029.

Jia, S. (Sixue) (2018) 'Behind the ratings: Text mining of restaurant customers' online reviews', *International Journal of Market Research*, 60(6), pp. 561–572. Available at: https://doi.org/10.1177/1470785317752048.

Poushneh, A. and Rajabi, R. (2022) 'Can reviews predict reviewers' numerical ratings? The underlying mechanisms of customers' decisions to rate products using Latent Dirichlet Allocation (LDA)', *Journal of Consumer Marketing*, 39(2), pp. 230–241. Available at: https://doi.org/10.1108/JCM-09-2020-4114.

Sánchez-Franco, M.J., Navarro-García, A. and Rondán-Cataluña, F.J. (2019) 'A naive Bayes

strategy for classifying customer satisfaction: A study based on online reviews of hospitality services', *Journal of Business Research*, 101, pp. 499–506. Available at: https://doi.org/10.1016/j.jbusres.2018.12.051.

Schoenmueller, V., Netzer, O. and Stahl, F. (2020) 'The Polarity of Online Reviews: Prevalence, Drivers and Implications', *Journal of Marketing Research*, 57(5), pp. 853–877. Available at: https://doi.org/10.1177/0022243720941832.

Zahoor, K., Bawany, N.Z. and Hamid, S. (2020) 'Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning', in *2020 21st International Arab Conference on Information Technology (ACIT)*. *2020 21st International Arab Conference on Information Technology (ACIT)*, pp. 1–6. Available at: https://doi.org/10.1109/ACIT50332.2020.9300098.