Application of Machine Learning to Mining Customer Reviews

Completed Research

Amir Abbas Darbanibasmanj
Telfer School of Management
University of Ottawa
amirabbas.darbani@gmail.com

Ajax Persaud
Telfer School of Management
University of Ottawa
ajax@telfer.uottawa.ca

Umar Ruhi
Telfer School of Management
University of Ottawa
Umar.Ruhi@uottawa.ca

Abstract

Online customer reviews are important sources of information influencing consumers' attitudes towards products and brands. Businesses also use them to gain deeper insights into consumers' perceptions, attitudes, and behaviors. This study uses machine learning (ML) and participant-based attributes of reviewers to classify them into distinct segments. The segments are then labelled and used to build a predictive model of customer behavior, which can help companies quickly profile customers and develop appropriate marketing strategies. The results show that our machine learning approach coupled with participation-based attributes generated unique clusters that are consistent with prior classification of online audiences. The personas-based clusters can help marketers make better use of reviewers in marketing campaigns by engaging them differently based on their interests and status in the online community. This study opens the door for further research using ML with larger and different review sites coupled with additional psychological, social, and economic variables.

Keywords

Machine learning, online customer reviews, customer attributes, Yelp, restaurants.

Introduction

Online customer reviews (OCR) are important sources of information shaping consumers' buying decisions (Castelli et al. 2017). Similarly, OCR also enables businesses to gain deeper insights into consumers' attitudes and behaviors towards products and brands (Abrahams et al. 2012; Bollen et al. 2011). However, their massive volume and often conflicting nature make it challenging to generate valuable strategic insights. Extant studies have used a variety of supervised ML techniques such as classification, text mining, and sentiment analysis to extract insights from OCR (e.g., Bollen et al. 2011; Castelli et al. 2017; Coussement et al. 2015; Fernández-Gavilanes et al. 2016; Finch 1999; Giatsoglou et al. 2017; Kang and Park 2014; Khashei and Bijari 2010; Law et al. 2017). There is a dearth of studies employing unsupervised ML techniques (Chaovalit and Zhou 2005) and behavioral data to gain deeper insights from OCR.

This study uses unsupervised ML and behavioral data to classify reviewers and build a predictive model of customer behavior. We choose unsupervised ML because while it is marginally less accurate than supervised ML, it is more efficient to use in real-time applications (Chaovalit & Zhou, 2005) and thus enables automatic mining of OCR, which can significantly improve timeliness of usage of OCR by companies. The goal of this study is to discover and validate personas of online reviewers based on social and informational attributes of their contributions and interactions. Segmentation of online reviewers can help online platforms hone in on different types of encouragement and motivation to facilitate greater reviewer participation, and also provide targeted advertising and sales promotion opportunities to businesses. Its practical focus is designed to show the relative ease with which firms and online platforms such as Yelp can mine OCR in real tie. For this study, we crawled 3,400 reviews of restaurants along with the reviewers' information and attributes from Yelp.com. Sussman and Siegal (2003) showed that reviewers' information and attributes foster a better understanding of the reviews and positively affect customers trust in the reviews. Based on prior research, we focused on three attributes: (1) size of the network of the reviewer in terms of the number of friends (Cheng and Ho 2015), the number of posted reviews (Hu et al. 2008), and the number of image posts (Cheng and Ho 2015).

The remainder of this paper is structured as follows. The next section provides a review of the relevant literature on OCR and ML techniques. Section three describes the methodology and Section four presents the results. Section five presents the discussion and implications.

Literature Review

Online reviews and engagement

Online reviews have become a widely accepted medium for customers to express their opinions about purchases and brands (Tang et al. 2009). Online reviews take many forms such as numerical rating scales, thumbs up/down, like/dislike, stars, and textual descriptions. Mudambi and Schuff (2010) showed that review extremity, review depth, and product type affect the perceived helpfulness of reviews. Specifically, they found that word count, review ranking, and product attributes affect readers' attitude towards the usefulness of reviews. Generally, reviews with higher word counts are perceived to be more useful to readers as they represent greater depth of information. Source credibility and argument quality are also shown to influence customers' attitudes towards reviews usefulness (Sussman and Siegal 2003) since they increase readers' trust in the source of the information (Chaiken 1980; Cheng and Ho 2015).

In addition to the attributes of the review itself, other researchers focus on the characteristics of the reviewers (Forman et al. 2008; Hu et al. 2008; Li and Hitt 2008; Mudambi and Schuff 2010; Smith et al. 2005). Source credibility, trustworthiness and expertise can affect perceptions toward the message and the perceived source of information affects readers' evaluation of the reviewer (Dou et al. 2012; Lee and Choeh 2017). According to Cheng and Ho (2015), readers perceive reviews as more helpful when the reviewers' level of socialization was higher i.e. the more followers a reviewer has, the greater the number of people who knows the person, thus the more useful the review is perceived by readers. They also found that a review is perceived as more useful when accompanied by images - image counts related to the reviews is an important determinant of reviews' perceived quality (Cheng and Ho 2015). Moreover, Hu et al. (2008) showed that readers pay attention to reviewer's reputation and exposure. Generally, consumers may pay more attention to higher exposure reviewers i.e. consumers respond better to reviews written by reviewers with higher exposure and reputation. Exposure is measured by the number of times a reviewer writes reviews (Hu et al. 2008).

In this study, the number of friends (level of socialization), image posts, and number of reviews (Cheng and Ho 2015; Hu et al. 2008) are the main determinants of reviewers' perceived helpfulness and credibility, and as the attributes for clustering reviewers based on their level of online engagement.

In terms of clustering reviewers based on their online engagement, Blanchard and Markus (2004) divided online communities into three different personas: Leaders, Participants, and Lurkers. "Leaders" are influential users who are active in the online environment and are generally well-respected. "Participants" are users who engage by posting messages but are not considered leaders. "Lurkers" are people who are the least engaged users - they read but rarely post (Blanchard and Markus 2004). In another study, Kozinets (1999) classified online social communities using a two-dimensional approach: Consumption Activity and Ties to the Community. Based on these two dimensions, they identified four categories of online community personas – *Tourists, Minglers, Devotees,* and *Insiders*. "Tourists" are members who do not have strong social ties to the community and their consumption activity interests are superficial. "Minglers" have strong social ties but are not very interested in the central consumption activity. "Devotees" are quite interested in the consumption activity but have fewer social ties to the community. "Insiders" have strong social attachment to the community and to the consumption activity (Kozinets 1999).

Machine learning methods (clustering and classification)

Clustering techniques partition a sample into different groups so that the objects in a group are similar to each other within the cluster while dissimilar to objects in other clusters (Han et al. 2011; Larose and Larose 2014). Clustering is also known as unsupervised classification as there are no predefined classes in the dataset. Clustering techniques have been widely used for customer segmentation. Clustering algorithms can be classified into partitioning, hierarchical, density-based, grid-based, model-based, frequent pattern-based, and constraint-based approaches (Han et al. 2011). Six of the frequently used clustering algorithms are k-means, expectation–maximization (EM), COBWEB, repeated-bisection approach, graph-partitioning algorithm, and density-based method (Kou et al. 2014). K-means is one of the most popular algorithms for clustering. This algorithm uses the mean value in order to assign each item to the cluster (MacQueen 1967). The main advantage of this algorithm is that its complexity is linear and its execution time depends on the number of samples and as a result it can be used with large datasets (Tufféry 2011).

Our study uses K-means. The process for K-means is presented as follows (Pandya and Macy 1995):
Step 1. Initialize:

Choose the number of clusters, k. For each of these k clusters chooses an initial cluster center: $\{c_1(m), c_2(m), \ldots, c_k(m)\}$, where $c_j(m)$ represents the value of the cluster center at the mth iteration.

Step 2. Distribute samples:

Distribute all sample vectors.
$x_p \in \vartheta_j(m)$
If $\left\|x_p - c_j(m)\right\| < \left\|x_p - c_i(m)\right\|$ for all $i = 1,2,\ldots,k, i \neq j,$
$\vartheta_j(m)$ Represents the population of cluster j at iteration m.

Step 3. Calculate new cluster centers:

$$c_j(m+1) = \frac{1}{M} \sum_{x_p \in \vartheta_j(m)} x_p$$

where M_j is the number of sample vectors attached to ϑ_j during Step 2.

Step 4. Check for convergence:
The condition for convergence is that no cluster center has changed its position during Step 3.

The purpose of classifying data is to convert it to meaningful homogeneous classes and use it for future predictions. Data classification is used in many industries such as healthcare, banking, advertising, and consumer packaged goods. In the context of ML, classification is made in terms of four different tasks: binary, multi-class, multi-labelled, and hierarchical. Sokolova and Lapalme (2009) present a systematic analysis of twenty-four performance measures used in these tasks.

Binary classification is one of the most popular classification tasks where the input is to be classified into one, and only one, of two non-overlapping classes which are typically named as positive and negative classes. The four counts which constitute a confusion matrix (Table 1) for binary classification are: the number of correctly recognized positive class examples (true positives: TP), the number of correctly recognized examples that belong to the negative class (true negatives: TN), and cases that are either incorrectly assigned to the positive class (false positives: FP) or that are not recognized as positive class examples (false negatives: FN) (Sokolova and Lapalme 2009).

Predicted Class		**Actual Class**	
		Positive	*Negative*
	Positive	TP	FP
	Negative	FN	TN

Table 1. Binary Confusion Matrix

Precision, Recall and Accuracy are some of the performance measures used for binary classification and are defined as follows (Duman et al. 2012):

Precision= TP/(TP+FP)
Recall= TP / (TP+FN)
Accuracy= (TP+TN)/(TP+FN+FP+TN)

Another measure to calculate the accuracy of the classification (AUC) models is area under the receiver operating characteristics (ROC) curve. Hand and Anagnostopoulos (2013) state that the area under the ROC curve (AUC) is a very widely used measure of performance for classification and diagnostic rules. The AUC evaluates different classifiers using different metrics. Waegeman et al. (2008) reveals that the area under the receiver operating characteristics curve is increasingly used as a performance measure for binary classification systems.

There are several classification algorithms developed by scholars and researchers. Six of the most frequently used ones are C4.5, Neural Networks, Classification and Regression Tree (CART), Support Vector Machine (SVM), Naïve Bayes and Logistic Regression (Peng et al. 2011).

Methodology

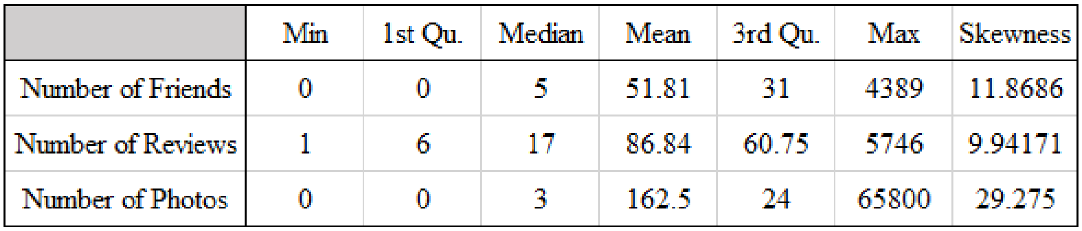
In our study, we crawled from Yelp's website the customer reviews and their authors' attributes of the top 40 reviewed restaurants of the City of Ottawa, Canada. The total number of reviews crawled is 3,400. The three main attributes extracted and used for further clustering and persona identification analysis are: size of the network of the author in terms of number of friends (Cheng and Ho 2015), number of posted reviews (Hu et al. 2008), and number of image posts (Cheng and Ho 2015).

After selecting the relevant features, we cleaned our crawled dataset and prepared it for further analysis and the remaining steps. In the third step, preparing the data for the clustering step (and to prevent overfitting), the top 95 percentile of the reviewers (based on the selected attributes) were removed from the data and had set aside to be analyzed in the last step (as the outlier cluster). In the normalization step, we scaled the three attributes (number of friends, number of reviews, and number of images posts) to be passed to next step, clustering of the data.

Following this, we used one of the most widely-used clustering algorithms, K-means, to cluster the data into several different clusters. To find the optimum number of clusters (from a ML perspective), we applied a well-known clustering evaluation model, the Silhouette Coefficient (Tan 2018). In the next step, selecting best k, combining the Silhouette coefficient results and related academic research on online and digital personas, we chose the optimum number of clusters. We then labeled each of the cluster by matching our clusters' characteristics with the online community persona characteristics and nomenclature proposed by Kozinets (1999) i.e. Tourists, Minglers, Devotees, and Insiders.

Analysis & Results

For this study, using R programming language and the crawled data of 3,400 reviews, we clustered the authors (reviewers) based on three different attributes: number of reviews, number of image posts, and number of friends. The brief summary of the attributes in tabular and graphical form are presented in Table 2 and Figures 1, 2, 3, and 4 below.



	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Skewness
Number of Friends	0	0	5	51.81	31	4389	11.8686
Number of Reviews	1	6	17	86.84	60.75	5746	9.94171
Number of Photos	0	0	3	162.5	24	65800	29.275

Table 2. Summary Descriptive Attributes of Reviewers

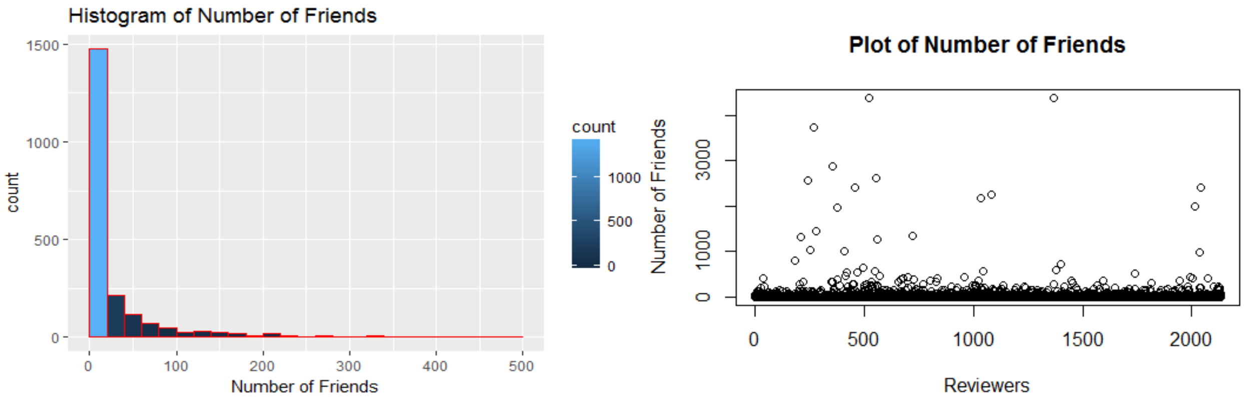


Figure 1. Graphical Representation of Number of Friends

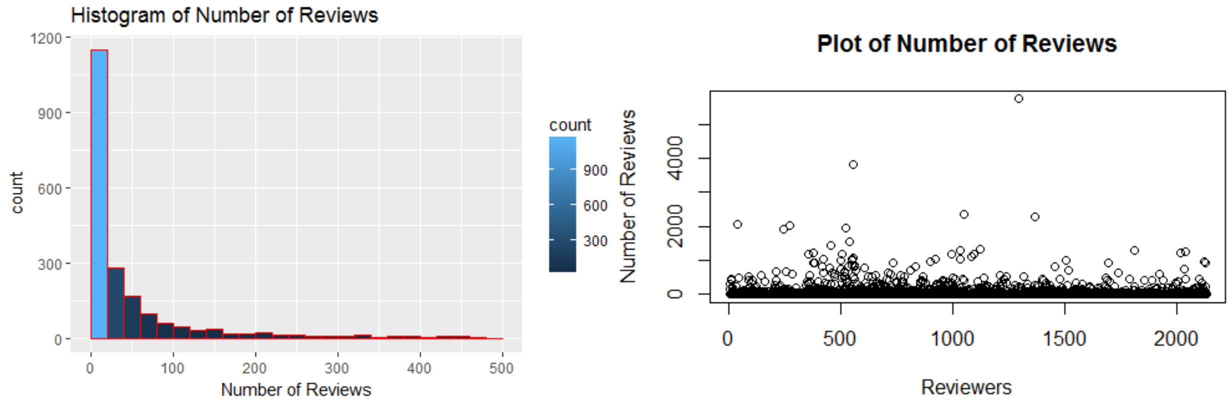


Figure 2. Graphical Representation of Number of Reviews

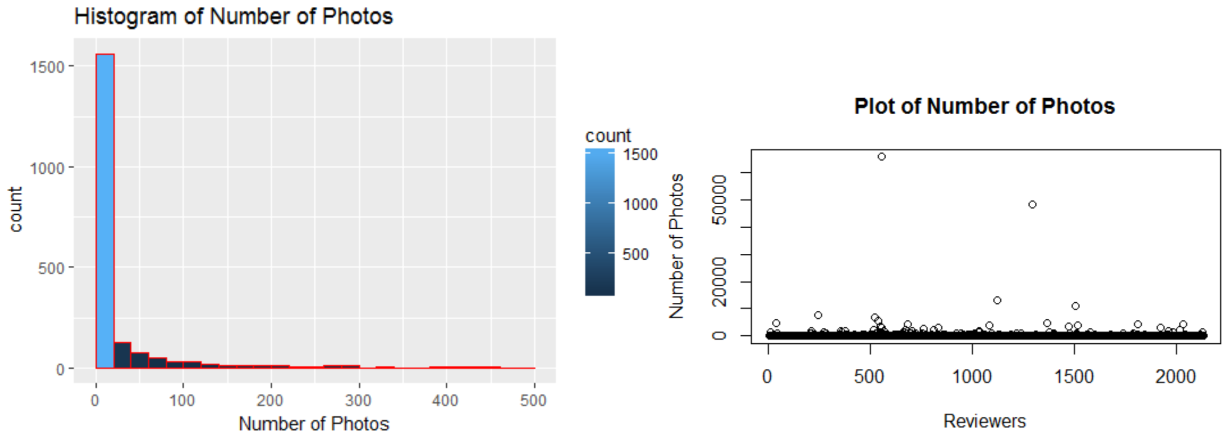


Figure 3. Graphical Representation of Number of Photos

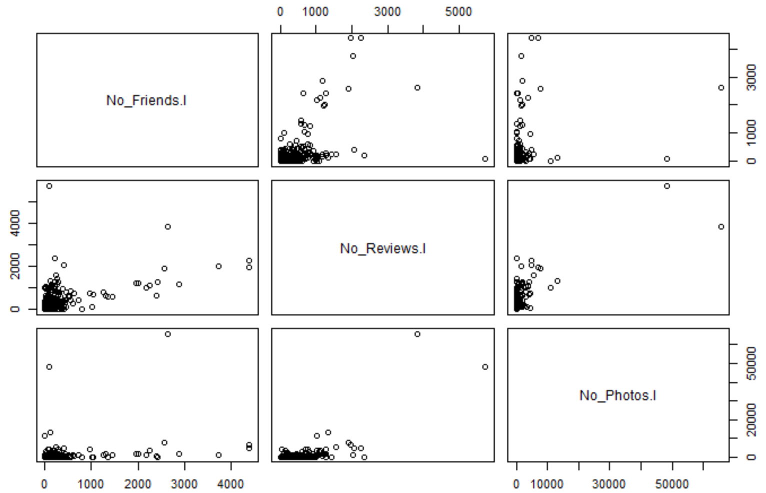
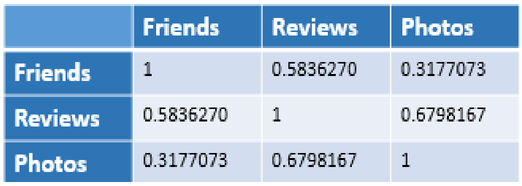


Figure 4. Summary of all Dimensions

The correlation matrix between the various attributes are displayed in Table 3.



	Friends	Reviews	Photos
Friends	1	0.5836270	0.3177073
Reviews	0.5836270	1	0.6798167
Photos	0.3177073	0.6798167	1

Table 3. Correlations Among Attributes

Before conducting the clustering, we identified the top 95 percentile reviewers as outliers (using the three features), and put the cluster of outliers aside, to be analyzed further after conducting the main clustering part. Then, using R Programming language and K-means clustering algorithm, we clustered the data into 3 to 9 clusters and measured the clustering quality using Silhouette index. The result of the clustering, 3 to 9, is displayed in Figure 5.

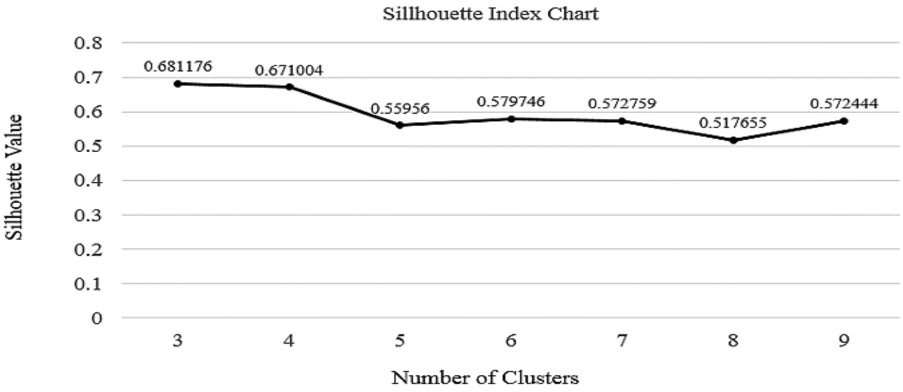


Figure 5. Silhouette Index

The Silhouette Index is a number between -1 and 1 and the greater the silhouette index (closer to 1), the better the clustering result (Tan 2018). As shown in the chart above, a cluster of 3 has the highest silhouette value (0.681176), thus the optimum number of clusters would be 3. However, considering the outliers as another extreme cluster, the total amount would be 4 clusters.

Reviewing the characteristics of the personas and our clusters separately, we observe a fairly good match between each cluster and the persona labels. In our schema, Cluster 1 corresponds to “Minglers”, cluster 2 to “Tourists”, cluster 3 to “Devotees”, and cluster 4 to “Insiders”. The clustering result along with the appropriate persona labels are displayed in Table 4. Figure 6 provides a visualization of the clusters.

Cluster	Number of Friends (Average)	Number of Reviews (Average)	Number of Photos (Average)	Population	F.R.P	Personas
Cl.1	73.23	135.5	42.4	220	HMM	Minglers
Cl.2	7.9	20.55	8.93	1634	LLL	Tourists
Cl.3	55.66	154.8	262.5	79	MMH	Devotees
Cl.4 (Outliers)	383	545.8	1503	201	(+H)(+H)(+H)	Insider

Table 4. Descriptive Statistics of Clusters

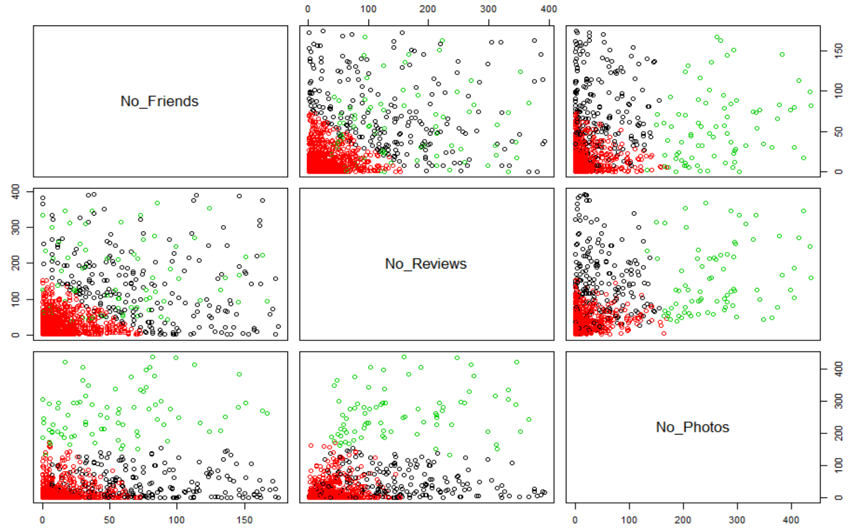


Figure 6. Cluster 1 to 3 Visualization

Predictive model

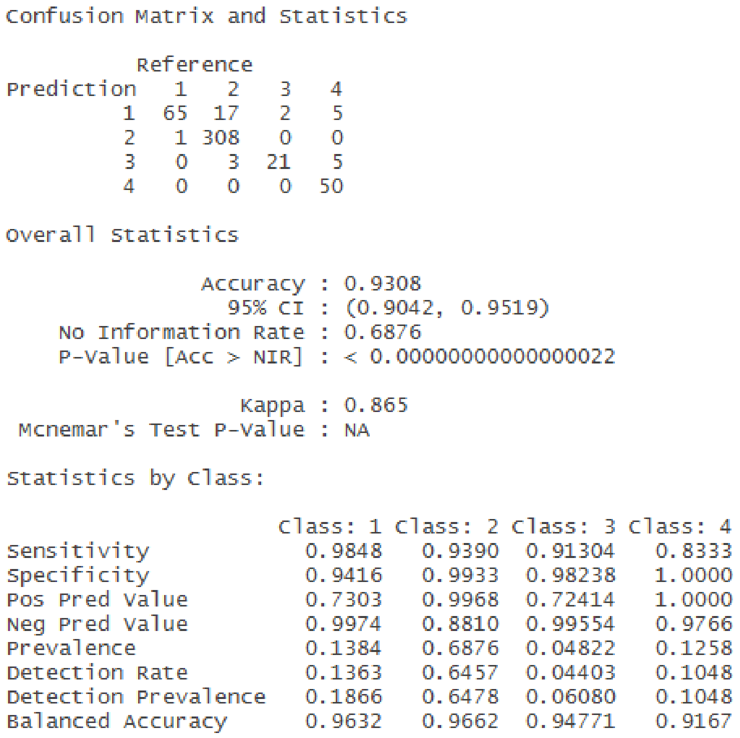
In order to build the predictive model, the attributes Number of Friends, Number of Reviews, Number of Photos along with two new attributes – Average Word Count of the textual reviews (length) and the Average Rating were used as predictors of the personas. For training the model, we used a k-fold cross-validation model with k equal to 10, with 3 repeats. It is also important to note that due to having imbalanced classes, we used CARET’s (an R package) over-sampling method, to increase the size of the minority class. The algorithms of the choices are CART, KNN, and SVM. For building the predictive model, R programming language and CARET package were used. The result of the predictive models is displayed in Table 5 below.

The results indicate that the KNN algorithm provides a better accuracy (93%), alongside better sensitivity and specificity from the other algorithms. Sensitivity and specificity are important because they show how the model is performing in predicting each of the classes uniquely.

Algorithm	Accuracy	Sensitivity				Specificity			
		Class.1	Class.2	Class.3	Class.4	Class.1	Class.2	Class.3	Class.4
CART	0.8218	0.8333	0.8476	0.9565	0.6166	0.8467	0.9060	0.9823	1.0000
KNN	0.9308	0.9848	0.9390	0.9130	0.8333	0.9416	0.9933	0.9823	1.0000
SVM	0.91	0.8681	0.9485	0.8775	0.7766	0.9490	0.9338	0.9790	0.9944

Table 5. Accuracy, Sensitivity, and Specificity of Predictive Model

The following is the outcome of the predictive model using the KNN algorithm:



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  65  17   2   5
         2   1 308   0   0
         3   0   3  21   5
         4   0   0   0  50

Overall Statistics

               Accuracy : 0.9308
                 95% CI : (0.9042, 0.9519)
    No Information Rate : 0.6876
    P-Value [Acc > NIR] : < 0.0000000000000022

                  Kappa : 0.865
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.9848   0.9390  0.91304   0.8333
Specificity            0.9416   0.9933  0.98238   1.0000
Pos Pred Value         0.7303   0.9968  0.72414   1.0000
Neg Pred Value         0.9974   0.8810  0.99554   0.9766
Prevalence             0.1384   0.6876  0.04822   0.1258
Detection Rate         0.1363   0.6457  0.04403   0.1048
Detection Prevalence   0.1866   0.6478  0.06080   0.1048
Balanced Accuracy      0.9632   0.9662  0.94771   0.9167
```

Discussion

In this study we looked at online customer review mining and insight analysis from a new methodological point of view and approach, using both supervised and unsupervised ML techniques. As far as we know, such techniques have not been used in analyzing online consumer reviews and user-generated content. Our proposed methods aim to complement those used in other studies which primarily focus on text mining of online reviews through feature extraction, content analysis, and sentiment analysis. By using participation-based attributes of reviewers such as their number of friends in the online community, the number of reviews they write, and the number of photos posted by them, we propose a method to further understand the social and informational aspects of user-generated content on consumer review portals.

By using supervised and unsupervised ML techniques in tandem, we show that the source of the reviews (reviewers) are not homogenous, and their social and informational attributes form the basis of unique personas. To interpret these personas, we adopt Kozinets' (1999) widely used taxonomy of online community members. The alignment of our resulting clusters offers a validation of Kozinets' taxonomy of online community members. *Tourists* are people who don't have strong social ties to the community and their level of activity is low – hence assigned to cluster 2; *Minglers* have strong social ties to the community but are not necessarily highly active (cluster 1); *Devotees* are quite active but exhibit fewer social ties (cluster 3); and *Insiders* are the group that has significant social ties and are also highly active (cluster 4). To further validate these clusters, we used a supervised (classification) model which can predict the persona of the reviewers based on attributes of word count of reviews, and the average ratings output of each cluster. Using these attributes, our model demonstrates a high level of predictive accuracy of our clusters.

This research has several implications for practice. For online consumer review platforms such as Yelp, reviewer personas can help delineate different types of participation behaviors, which can be used by these platforms to develop strategies to encourage greater participation. For example, recommendations for like-minded participants with similar interests can be provided for reviewers in the *Devotees* cluster, while *Minglers* can be stimulated to post additional content through highlighting positive ratings on past reviews. Segmentation of online reviewers in such a fashion would also allow the review platform to offer targeted advertising opportunities to sponsors who might want to focus on specific types of reviewers according to their social or informational footprints. Typically, such targeting is only based on demographic attributes of community members, and not on persona characteristics. For companies that are being reviewed (e.g. restaurants), reviewer personas can provide a potential opportunity to identify influencers and better engage them during promotional campaigns. Through targeted sales promotions, *Devotees* can be engaged to try new offerings and generate fresh online content, while *Minglers* and *Insiders* with high reach can be engaged to spread the word about product offerings to their networks of friends and followers.

Conclusion

This study was undertaken to investigate whether ML combined with participation-based attributes of reviewers such as their number of friends in the online community, the number of reviews they write and the number of photos they post are effective alternative to current methods for review mining and insight analysis. We observe that this approach is not yet widely used in review mining and thus represents an early attempt at exploring this approach. Using online review data from Yelp on restaurants located in a major Canadian city, we observe that our ML approach coupled with participation-based attributes were valid in the sense that they generated unique clusters that are not only consistent with prior classification of online audiences but generated insights that can inform marketing strategies and campaigns. Essentially, the characteristics of the four unique clusters or personas generated – Tourists, Minglers, Devotees and Insiders - can help online review platforms like Yelp to better target advertisers by providing them more targeted information based on personas rather than demographic characteristics. Additionally, marketers can make better use of their reviewers in their marketing campaign by engaging them differently based on their interests and status in the online community. Finally, this study opens the door for additional research using ML with larger and different review sites and communities coupled with additional psychological, social, and economic variables.

Acknowledgements

This research was supported through a research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

REFERENCES

Abrahams, A. S., Jiao, J., Wang, G. A., and Fan, W. 2012. "Vehicle Defect Discovery from Social Media," *Decision Support Systems* (54:1), pp. 87-97.

Blanchard, A. L., and Markus, M. L. 2004. "The Experienced Sense of a Virtual Community: Characteristics and Processes," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* (35:1), pp. 64-79.

Bollen, J., Mao, H., and Zeng, X. 2011. "Twitter Mood Predicts the Stock Market," *Journal of computational science* (2:1), pp. 1-8.

Castelli, M., Manzoni, L., Vanneschi, L., and Popovič, A. 2017. "An Expert System for Extracting Knowledge from Customers' Reviews: The Case of Amazon. Com, Inc," *Expert Systems with Applications* (84), pp. 117-126.

Chaiken, S. 1980. "Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion," *Journal of personality and social psychology* (39:5), p. 752.

Chaovalit, P., and Zhou, L. 2005. "Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches," *Proceedings of the 38th annual Hawaii international conference on system sciences*: IEEE, pp. 112c-112c.

Cheng, Y.-H., and Ho, H.-Y. 2015. "Social Influence's Impact on Reader Perceptions of Online Reviews," *Journal of Business Research* (68:4), pp. 883-887.

Coussement, K., Benoit, D. F., and Antioco, M. 2015. "A Bayesian Approach for Incorporating Expert Opinions into Decision Support Systems: A Case Study of Online Consumer-Satisfaction Detection," *Decision Support Systems* (79), pp. 24-32.

Dou, X., Walden, J. A., Lee, S., and Lee, J. Y. 2012. "Does Source Matter? Examining Source Effects in Online Product Reviews," *Computers in Human Behavior* (28:5), pp. 1555-1563.

Duman, E., Ekinci, Y., and Tanrıverdi, A. 2012. "Comparing Alternative Classifiers for Database Marketing: The Case of Imbalanced Datasets," *Expert Systems with Applications* (39:1), pp. 48-53.

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., and González-Castaño, F. J. 2016. "Unsupervised Method for Sentiment Analysis in Online Texts," *Expert Systems with Applications* (58), pp. 57-75.

Finch, B. J. 1999. "Internet Discussions as a Source for Consumer Product Customer Involvement and Quality Information: An Exploratory Study," *Journal of Operations Management* (17:5), pp. 535-556.

Forman, C., Ghose, A., and Wiesenfeld, B. 2008. "Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information systems research* (19:3), pp. 291-313.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., and Chatzisavvas, K. C. 2017. "Sentiment Analysis Leveraging Emotions and Word Embeddings," *Expert Systems with Applications* (69), pp. 214-224.

Han, J., Pei, J., and Kamber, M. 2011. *Data Mining: Concepts and Techniques*. Elsevier.

Hand, D. J., and Anagnostopoulos, C. 2013. "When Is the Area under the Receiver Operating Characteristic Curve an Appropriate Measure of Classifier Performance?," *Pattern Recognition Letters* (34:5), pp. 492-495.

Hu, N., Liu, L., and Zhang, J. J. 2008. "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects," *Information Technology and management* (9:3), pp. 201-214.

Kang, D., and Park, Y. 2014. "Based Measurement of Customer Satisfaction in Mobile Service: Sentiment Analysis and Vikor Approach," *Expert Systems with Applications* (41:4), pp. 1041-1050.

Khashei, M., and Bijari, M. 2010. "An Artificial Neural Network (P, D, Q) Model for Timeseries Forecasting," *Expert Systems with applications* (37:1), pp. 479-489.

Kou, G., Peng, Y., and Wang, G. 2014. "Evaluation of Clustering Algorithms for Financial Risk Analysis Using Mcdm Methods," *Information Sciences* (275), pp. 1-12.

Kozinets, R. V. 1999. "E-Tribalized Marketing?: The Strategic Implications of Virtual Communities of Consumption," *European Management Journal* (17:3), pp. 252-264.

Larose, D. T., and Larose, C. D. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons.

Law, D., Gruss, R., and Abrahams, A. S. 2017. "Automated Defect Discovery for Dishwasher Appliances from Online Consumer Reviews," *Expert Systems with Applications* (67), pp. 84-94.

Lee, S., and Choeh, J. Y. 2017. "Exploring the Determinants of and Predicting the Helpfulness of Online User Reviews Using Decision Trees," *Management Decision* (55:4), pp. 681-700.

Li, X., and Hitt, L. M. 2008. "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research* (19:4), pp. 456-474.

MacQueen, J. 1967. "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*: Oakland, CA, USA, pp. 281-297.

Mudambi, S. M., and Schuff, D. 2010. "What Makes a Helpful Review? A Study of Customer Reviews on Amazon. Com," *MIS quarterly* (34:1), pp. 185-200.

Pandya, A. S., and Macy, R. B. 1995. *Pattern Recognition with Neural Networks in C++*. CRC press.

Peng, Y., Kou, G., Wang, G., and Shi, Y. 2011. "Famcdm: A Fusion Approach of Mcdm Methods to Rank Multiclass Classification Algorithms," *Omega* (39:6), pp. 677-689.

Smith, D., Menon, S., and Sivakumar, K. 2005. "Online Peer and Editorial Recommendations, Trust, and Choice in Virtual Markets," *Journal of interactive marketing* (19:3), pp. 15-37.

Sokolova, M., and Lapalme, G. 2009. "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management* (45:4), pp. 427-437.

Sussman, S. W., and Siegal, W. S. 2003. "Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption," *Information systems research* (14:1), pp. 47-65.

Tan, P.-N. 2018. *Introduction to Data Mining*. Pearson Education India.

Tang, H., Tan, S., and Cheng, X. 2009. "A Survey on Sentiment Detection of Reviews," *Expert Systems with Applications* (36:7), pp. 10760-10773.

Tufféry, S. 2011. *Data Mining and Statistics for Decision Making*. John Wiley & Sons.

Waegeman, W., De Baets, B., and Boullart, L. 2008. "Roc Analysis in Ordinal Regression Learning," *Pattern Recognition Letters* (29:1), pp. 1-9.