

In the Eye of the Reviewer: An Application of Unsupervised Clustering to User Generated Imagery in Online Reviews

Gijs Overgoor
University of Amsterdam
g.overgoor@uva.nl

Rohan Mestri
North Carolina State University
rsmestri@ncsu.edu

William Rand
North Carolina State University
wmrand@ncsu.edu

Abstract

Mining opinions from online reviews has been shown to be extremely valuable in the past decades. There has been a surge of research focused on understanding consumer brand perceptions from the textual content of online reviews using text mining methods. With the increase in smartphone usage and ease of posting images, these reviews now often contain visual content. We propose an unsupervised cluster method to understand the user-generated imagery (UGI) of online reviews in the travel industry. Using the deep embedded clustering model we group together similar UGI and examine the average review ratings of these clusters to identify imagery associated with positive and negative reviews. After training the method on the entire dataset, we map out individual hotels and their corresponding UGI to show how hotel managers can use the method to understand their performance in particular areas of customer service based on UGI. The performance in a cluster relative to the population can be a clear indicator of areas that need improvement or areas that should be highlighted in the hotel's marketing efforts. Overall, we present a useful application using visual analytics for mining consumer opinions and perceptions directly from image data.

1. Introduction

72% of consumers always or frequently read reviews before deciding where to visit, eat or stay, and on average those users read nine reviews before making a decision to book a hotel or a restaurant.¹ Online reviews have become an integral part of the online travel industry. Consumers search out reviews when making decisions and these reviews have an important impact on a consumer's decision to book a hotel or to visit a restaurant [1]. In turn, for businesses

they are a valuable source of feedback to understand customer experiences and perceptions [2]. Businesses can improve their services and facilities based on these insights [3]. Clearly, it's crucial for business to understand what the consumer is saying about their business. Generally, there are 3 types of information presented to a consumer in the form of reviews. Review score (numerical), review text (textual) and review imagery (visual). The numerical and textual parts of the review have been covered extensively in the past, but the visual component of reviews has not received sufficient attention.

Smartphones and the ease of capturing and sharing photos online have led to a great increase of UGI online. More often reviews are now accompanied with UGI. For example, there are over 160M photos generated by travelers on TripAdvisor.² Given the evidence that images are more engaging and hold more information than text [4], one can imagine that the UGI in online reviews holds valuable information for businesses. For instance, Zhang et al. [5] show that UGI on Yelp provides information about a restaurant's survival potential. Ma et al. [6] are one of the very few studies that investigate UGI in the online review context. They indicate that the main reason it is an understudied area is likely due to technical difficulties in translating the images into structured information.

The extraction of information from images, or image mining, has been shown to be useful in recent research. There are several marketing studies that now utilize these methods to connect imagery to interesting marketing problems [5, 7, 8, 9]. As for online reviews, Ma et al [10] establish that UGI contains useful information for predicting the helpfulness of reviews. Despite these developments, there is no research that has shown to be effective in opinion mining from UGI in online reviews.

We propose an unsupervised clustering method, based on the deep embedded clustering method [11], to

¹<https://tripadvisor.mediaroom.com/2019-07-16-Online-Reviews-Remain-a-Trusted-Source-of-Information-When-Booking-Trips-Reveals-New-Research>

²<http://ir.tripadvisor.com/static-files/6d4c71fd-3310-48c4-b4c5-d5ec04e69d5d>

cluster UGI based on visual similarity. This method is completely unsupervised, which means that we do not need to teach or train our model to recognize or predict specific labels. Instead our method automatically detects UGI clusters with similar visual properties. Our method provides several directly managerially relevant outputs: (1) we identify and cluster UGI that users generally post with their reviews across different hotels, (2) by highlighting the distribution of review scores across these clusters we show what users generally post when they are satisfied vs. dissatisfied, (3) when looking at individual hotels we highlight high and low performance areas making it easy to identify places for improvement.

2. Background

Previous research has established the impact of online reviews on the online travel industry using a variety of methods [12]. On comparison websites such as TripAdvisor and Yelp the review information is generally presented to consumers by the average score review score, scores on aspects such as cleanliness and location and by individual consumer reviews [13]. The average scores can easily be used in statistical models to understand their impact on demand or consumer choices, but the individual reviews consist of unstructured data such as text and images that need processing techniques to transform into useful information [14]. A survey study about information sources and their importance for online hotel bookings showed that in terms of the types of information, consumer review scores are perceived as the most important source, followed by hotel images and descriptive (textual) information about hotels [1].

On the most popular platforms there is an abundance of reviews available and it is impossible for consumers or even businesses to process all this information manually. As an example, in 2014 there were on average 165,000 new reviews each day on TripAdvisor alone.³ As mentioned earlier, a user only reads nine reviews on average, which means that users rely on review summaries and platform filtering to examine the information. For these reasons, we rely on machine learning methods to examine this unstructured information. The textual component of reviews has received quite some attention in the past decade. Using NLP, previous work has examined the sentiment of text and the over-arching topic of a review and then used that information to make predictions about the review score or the review helpfulness [15].

Text mining methods have been used in a variety of applications, either to understand the person or group of

people generating and/or receiving textual content [12]. These methods are very helpful to generate marketing insights from textual content to understand consumers and their perception of brands. Several (recent) studies have investigated the use of text mining to summarize a large number of reviews online, mostly with the intent of providing consumers searching for information with the most helpful content. Applications of text mining to reviews range from exploring customer satisfaction [16], identifying the most informative reviews and sentences on TripAdvisor [17] to mining consumer opinions and sentiments [18]. For example, Tsai et al. [15] first classify reviews as helpful vs. non-helpful and then highlight the hotel features in the helpful reviews. Based on these features and their helpfulness scores platforms could enhance search functions to allow consumers to filter or group based on what they are interested in. In general, research using text mining methods has highlighted how to extract information from reviews at scale and how this information can be used to improve platforms or identify performance issues or highlights for businesses. A major limitation is that these studies look only at the textual and/or numerical component of reviews and overlook the visual component that is often a part of consumers' reviews. Although, the textual component plays a prominent role in the review, in recent years the image has become increasingly helpful in reviews [10].

Images are now increasingly available, because of smartphones and online platforms that make it easy to upload an image along side text. People process visual information more easily than text [19] and we have seen ample evidence that content online that has a visual component is much more engaging [4]. For these reasons, it is crucial to understand what is presented in these images. With the refinement of image mining methods in the past decade, we are becoming increasingly capable of extracting useful information from imagery online. Images provide managers with a way to visually listen, i.e., better understand, to what consumers are posting about brands on social media [8]. Studies in the online travel industry also established the impact of hotel or AirBnB property images on consideration set formation and demand using image mining methods [7, 20, 9]. As for online reviews, Ma et al [10], propose a deep learning approach to predict the helpfulness of reviews and establish that UGI provides additional information about customer experiences and improve helpfulness prediction. There are currently no studies that offer a summary or overview of what the imagery in online reviews represent.

Most of the online review studies mentioned previously focus on providing a summary of useful

³<https://www.telegraph.co.uk/travel/lists/TripAdvisor-in-numbers/>

information that is presented in text-based reviews at a large scale. The main reason behind this focus on summaries is that there is simply too much information for consumers to process [15]. Another reason for the summarization is that it can detect underlying opinions in these reviews [17]. As a result, the methods turn out to be valuable to both the consumer and the business [2]. Images could play a similar role. A problem, however, is that images are very rich sources of information. A standard User-Generated Image on TripAdvisor has 224 by 224 pixels, or about 50k pixels each consisting of a Red, Green, and Blue channel. As a result, it is difficult to summarize what is portrayed by an image, let alone a set of images. For this reason, we turn to a very common method used for grouping data based on similarities: Clustering.

Clustering is a method of unsupervised learning (i.e., we do not need to give the machine feedback on what it is learning), which is a popular technique in machine learning to get a better understanding of our data. When we apply a clustering algorithm we group data points into groups and these groups can provide us with higher-level information about what the data looks like. For example, k-means clustering [21], can be used for customer segmentation based on purchase history, interests and demographics. In the online travel industry, Ahani et al. [22], show that by clustering consumers based on textual reviews and ratings they can understand why different travelers select certain spa hotels and how these hotels can use this information to better service their (potential) customers. The goal in our research is to understand consumers through the images they post with their online reviews.

Images are a very high-dimensional information source that needs additional processing. First, we need to translate the pixels of image using a Convolutional Neural Network (CNN) architecture [23]. A CNN uses convolutions to examine pictures through "filters". These filters are responsible for detecting certain patterns, where early layers detect simple image information (e.g., colors, lines, or edges) and later layer detect more complex image information (e.g., complex shapes, buildings, faces, etc.). Generally, the height and width dimension of the convolutional layers in an architecture such as the VGG16 [24] increases as the information is processed by new layers and the number of filters increases. Basically, this means that what the net is detecting gets more complex. Eventually, using other layers such as flattening and fully connected layers, an image is embedded onto a vector space. The image is now translated from unstructured image information into structured information, or features, that can be input into a model. Our method uses the Deep

Embedded Clustering [11], that is shown to be much more effective for clustering imagery than standard clustering methods, such as k-means or hierarchical clustering. General clustering methods fail at clustering images, because of the high dimensions of the data, even if they were to be embedded on to a vector space before they are used in these methods. In the next section, we discuss the clustering method in more detail and then we highlight how we can use it to understand what consumers usually capture in their UGI and what we can do with this information.

3. Method

In unsupervised clustering, traditional algorithms like the k-Means algorithm generally fail on higher dimensional data and they are too computationally expensive. Deep Neural Network-based clustering methods have risen to prominence in the past few years to solve both these issues. We use the Deep Embedding Clustering model, first proposed in [11], to perform unsupervised clustering over high dimensional data. One of the primary advantages of this method is that, after training, we can still feed unseen samples into the model and the model maps the unseen samples to their most probable cluster. This is helpful, because we can cluster all reviews of a single hotel and then highlight the clusters associated with high or low review scores, to identify performance areas. At the same time, we can also directly label a new image as belonging to a high or low performance cluster, which is useful for managers to identify potentially effective marketing assets.

3.1. Deep Embedded Clustering

Stochastic Neighbor Embedding (SNE) [25] is a method proposed to reduce the dimensionality of a set of data points from a higher dimensional space to a lower dimensional space, while maintaining the neighborhood similarities in the lower dimensional space as observed in the higher dimensions. The neighborhood similarities are measured by a Gaussian Similarity Kernel function, both in the high dimensional space as well as in the lower dimensional space, and the difference between these similarities is minimized using the Kullback–Leibler (KL) divergence metric. In order to solve the crowding problem (i.e. limited space for "neighbors" in high-dimensional data when forced onto a 2-dimensional plane) observed in the SNE method, van der Maaten et al. [26] proposed to use the t-distribution based similarity kernel in the lower dimensional space and thus, formed the t-Distribution Stochastic Neighbor Embedding (tSNE).

Motivated by the tSNE, the Deep Embedded

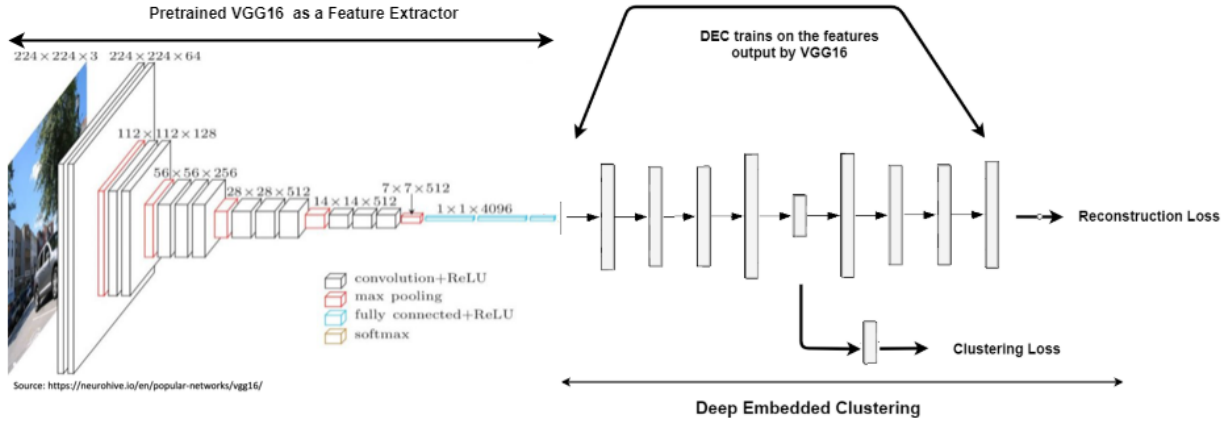


Figure 1. The above diagram shows the combined model of the Deep Embedded Clustering along with the VGG16 model at the preprocessing stage. The DEC model clearly shows the stacked denoising autoencoder with each of the vertical bars representing fully connected layers of proportional node size. The DEC model carries out the process of clustering these features into a discrete set. These clusters can be obtained from the latent space after detaching the decoder.

Clustering (DEC) model was developed using an autoencoder framework for dimensionality reduction. A Stacked Denoising Autoencoder learns the distribution of the input data, by training it to recreate the input data itself. The input X is fed to the encoder part of the autoencoder, which tries to embed the higher dimensional data X into the lower dimensional data z , by shrinking the data to pass through a bottleneck. The z -points in the latent space are then passed through the decoder, which then tries to reconstruct the initial input. This autoencoding is useful, because it forces the model to preserve as much information in the dimension reduction, otherwise the decoder wouldn't be able to reconstruct the input. This makes the latent space as informative of the imagery as possible, which makes the cluster assignment more effective as a result.

After training the autoencoder, we obtain a latent space from which we can obtain reasonable estimates of the initial cluster centres in the data distribution. The k -means algorithm then processes the latent space to find the initial cluster centers of the N_c clusters. Once, we have these cluster centers, we find the t -distribution based similarity from each embedded datapoint z to these N_c cluster centres. These similarities are computed in a probabilistic manner, indicating the probability that the datapoint z_i will lie in a cluster with mean μ_j (with degrees of freedom df and k iterating over each cluster) and is given as:

$$q_{ij} = \frac{(1 + \frac{\|z_i - \mu_j\|^2}{df})^{-\frac{df+1}{2}}}{\sum_k (1 + \frac{\|z_i - \mu_k\|^2}{df})^{-\frac{df+1}{2}}}$$

In the finetuning process, this probability distribution is then self-trained to follow a target

distribution, and the difference of these distributions is minimized using the KL Divergence metric. The target distribution is chosen in a way that it sharpens the probabilities of membership into a particular cluster, thus refining these clusters close to a single convergence point. This target distribution is given as:

$$p_{ij} = \frac{\sum_i \frac{q_{ij}^2}{q_{ij}}}{\sum_j \sum_i \frac{q_{ij}^2}{q_{ij}}}$$

Guo et al. [27] proposed an improved version of the DEC which also minimizes the reconstruction loss along with the clustering loss, as a dual loss function. This is shown to maintain the local structure preservation property. Similarly, in our paper, we have used this dual loss function for training the model. The autoencoder part of the DEC model tries to reconstruct the features given by the VGG-16 model. The reconstruction loss is a mean squared error loss between the actual features and the predicted features.

3.2. Transfer Learning

In [28], Guo et al explored the possibility of using convolutional layers in both the encoder as well as decoder to reconstruct the image. This is essential as compared to using the traditional fully connected autoencoder as the latter trains on direct pixel values and hence fails to capture the features provided by the convolutional layers. Similarly, we train on features as opposed to direct pixel values, but instead of using convolutional layers in the autoencoder, we use a pretrained CNN to transform the images into features.

We are using the VGG-16 model [29] which is pretrained on the Places dataset with 365 scene categories [29]. This network is effective at detecting common places, such as hotel rooms, pools, or parks among others. Previous research on the impact of imagery on online hotel bookings show that this pretrained CNN is an effective method for extracting features that can be used for click-through rate prediction [9]. Each image is fed into this pretrained model first and it outputs a vector of size 365, with each of the 365 nodes consisting of a probability of belonging to one particular class. The magnitude of the remaining nodes is largely diminished due to the usage of a softmax activation function and we can not use this 365-length vector as features. Hence, we replace the last layer of the pretrained model with a sigmoid activation function, which serve as reasonable estimates of the features. These features are then fed to the encoder. The decoder is used to reconstruct these features from the latent space and the reconstruction loss is used to train the encoder so it preserves as much of the information while shrinking the dimensions.

4. Results

In this section we describe the application of the DEC to a set of online reviews with UGI scraped from TripAdvisor. First, we describe the data, then we show the results of the clustering method to the entire dataset to understand the distribution of UGI across clusters. We then highlight three example hotels and what we can learn from the clustering of the UGI. And finally, we discuss how we can use the method to identify useful marketing assets.

4.1. Data

We test our method on a collection of reviews with images from a group of New York City hotels from TripAdvisor. In total, we collected 5499 online reviews resulting in 9155 UGI. The average review rating is 4.48/5 with a standard deviation of 0.92. About 60% of the reviews have a 5 star rating. This is expected as TripAdvisor highlights that about 87% of customers write a review about a positive experience.⁴ About 75% of the reviews with images have only a single image and more than 95% of these reviews have less than 5 images. The maximum number of images belonging to a single review is 28. We do not observe a significant correlation between the number of images and the review rating.

⁴<https://tripadvisor.mediaroom.com/2019-07-16-Online-Reviews-Remain-a-Trusted-Source-of-Information-When-Booking-Trips-Reveals-New-Research>

4.2. Overall clustering

We start with a clustering all images across all hotels. We feed the entire data into our clustering model without attaching any labels to these data samples. After obtaining the clusters, we associate each embedded datapoint with its numerical rating, since some reviews feature multiple images that means that every image associated with that review receives the same numerical rating.⁵ Given that δ_{ij} is the binary membership of data sample i in the cluster j and τ_i is the rating given by the user, to which the sample belonged. The aggregate rating of the cluster j can be calculated as

$$\rho_j = \frac{\sum_{i=1}^N \delta_{ij} \tau_i}{\sum_{i=1}^N \delta_{ij}}$$

After some exploration, we have set the number of clusters and the latent space dimension to 10. The resulting clusters are both specific and substantial, which is the most important trade-off to consider here. In the near future, we aim to do a more extensive hyperparameter tuning.

In Table 1, we have listed the ten (10) clusters with four (4) sample images that best represent each cluster. Note that the method is completely unsupervised, therefore it does not get any feedback on similarities between images from labels. Instead it is based solely on the imagery themselves. We observe that the images within each cluster are remarkably similar in terms of the content they depict. From the sample imagery we observe the following about the clusters:

- Cluster 1 - zoomed in images of specific details
- Cluster 2 - lobby, bar or general hotel area
- Cluster 3 - Seating areas within the hotels
- Cluster 4 - Views from the hotels
- Cluster 5 - Front or outside of the hotels
- Cluster 6 - Food and Drinks
- Cluster 7 - Hotel rooms
- Cluster 8 - Style-details
- Cluster 9 - Bathroom
- Cluster 10 - Empire State Building

This is the UGI that travelers to New York City share in their online reviews about hotels. When we look at the mean and standard deviation in particular, we

⁵The results are robust to a weighting based on the number of UGI per review, where a single image receives a higher weighting than images that belong to reviews with multiple images.

observe that Cluster 1 represents UGI most related to dissatisfaction (lowest average score) with the hotels, whereas Cluster 10 represents UGI most closely related to satisfaction (highest average score). The cluster with the lowest average score (1) shows UGI portraying aspects of the hotel experience that are subpar, such as close ups of a bed, a closet or a bathtub/shower. The highest rated cluster (10), in contrast, mainly shows the Empire State Building, which could be thought of as UGI portraying the experience. We also observe that the standard deviation for Cluster 1 is largest, meaning that there is a larger spread in satisfaction across this cluster. We also observe this for the bathroom cluster (9). The other 8 clusters have fairly comparable average ratings and standard deviations, even though they portray very different UGI. In the individual level analysis we can think of these clusters as performance areas, which encapsulate more than just the physical areas, but also the features, experiences, and stylistic elements of individual hotels.

4.3. Individual Hotels and Distribution across clusters

After training the model with the entire set of data, we then segment the data and bin them into groups, with each group corresponding to images belonging to one particular hotel. We are able to observe the cluster distribution per hotel. We can then compare the means and standard deviation of individual hotels to those of the population. This provides indicators for high or low performance areas for individual hotels. To illustrate this application we have selected three hotels. Hotel A, whose average ratings are below average, Hotel B, whose average ratings are similar to the population average, and Hotel C, whose average ratings are above average. Figure 5, shows the performance of the hotels in each of the 10 performance areas, represented by the clusters, in comparison to the population average for these performance areas. The green rectangles indicate where the hotels over perform (i.e., a hotel average rating that is significantly higher than the average of the other hotels) and the red box indicates where the hotel under performs (i.e., a hotel average rating that is significantly lower than the average of all hotels). We observe that the Hotel A under performs in three clusters: Cluster 2, Cluster 8 and Cluster 10. Recall, that these clusters represent the lobby/bar areas (2), style details (8), and Empire State Building pictures (10) respectively. This hotel is generally under performing, as reflected by a below average rating, but performs especially poorly in these three areas. Though it performs poorly overall, it does seem to have little issue

in Cluster 1, which means that in general it has very little problems with specific tangibles, such as cleanliness, messiness, or damage, which is what Cluster 1 typically identifies. This is also reflected by a very low standard deviation in this cluster as compared to the population. The manager of this hotel might not be able to do much about the view it has of the Empire State Building, but it could potentially address the lobby / bar issues and the style details. On the other hand, Hotel B, is an average hotel, which is reflected by most performance areas, though it does perform below average for the first cluster. A clear insight for the manager of Hotel B is to examine the images related to Cluster 1 and see what can be done to improve performance in this area. Hotel B over performs in Cluster 5, which generally represents the outside or front of the hotel. These might be architectural style indicators, which seem to be appreciated by the visitors of Hotel B, and the manager could highlight these in their marketing efforts. Finally, Hotel C is an above average hotel. We can observe that this hotel over performs consistently (also reflected by low standard deviation across clusters), and does exceptionally well in areas related to Clusters 1, 8, and 9. All these are related to the hotels interior. Hotel C performs very well overall, but especially the bathroom and style details are very well received by its patrons. A manager of this hotel could play up these details in marketing materials and on their website.

5. Discussion

Methods for understanding large scale unstructured data such as image mining or visual analytics are becoming increasingly important and useful. In the past decade we have seen a surge of studies on online reviews focused on opinion mining and summarization of textual content, but little research on visual content. Effectively summarizing UGI is difficult because of the high-dimensional nature of the images. In this research, we present an image mining method to understand what consumers portray in their UGI in online reviews. We leverage the information portrayed by UGI using Deep Embedded Clustering, a high-dimensional clustering method that is much more effective than traditional clustering methods, such as k-means. We apply transfer learning to a CNN originally trained to recognize 365 places to embed imagery onto a vector space and use these features to effectively cluster UGI from TripAdvisor reviews of hotels New York City. It is important to emphasize that the method is unsupervised, so it does not need any human feedback for training. The system automatically identifies the 10 clusters. The method can then stay "up-to-date" every time it is

fine-tuned on new data, but it can also directly distribute new imagery across the clusters to identify marketing stimuli.

In addition to the real-time managerial insights such a system generates, it also helps us develop a deeper understanding of consumer behavior with respect to UGI. Our results clearly show 10 main types of imagery that are generally posted by users. Using the distribution of UGI and corresponding reviews across these clusters we highlight that, in general, dissatisfied customers post zoomed in pictures of tangibles in the hotel, such as style features, or furnishing, cleanliness and damages. The satisfied customer, for our New York City data, tends to share the Empire State Building. This is in line with previous research on textual reviews [16]. The standard deviations of the clusters also provide us with information on the volatility of certain areas as compared to others. In general, the clusters portray some clear performance areas for hotels, not reflected by the ratings generally offered by website such as Yelp and TripAdvisor. This highlights a clear advantage of our method, but also lends itself for an interesting study on the performance of individual hotels or competition between hotels.

The application of our method on the three individual hotels show how managers can easily identify areas in which the hotel is over or under performing. We saw, for example, that Hotel A was performing very well in Cluster 1, even though it was under performing overall. It also highlighted some important areas where it was underachieving. Using this information a hotel manager can identify areas that might need work, but it can also use UGI from high performance clusters for their marketing assets. We saw that Hotel B, an average hotel in terms of rating, was clearly under performing in a basic service area such as Cluster 1. At the same time it showed that customers generally liked the look of the hotel. As for the high performance hotel, we observed that it was performing very consistently across the board, but even then we were able to highlight some areas the hotel might want to focus on in its marketing materials.

There are a number of areas where this method can be improved. We have limited ourselves to a single city and a group of hotels within that city. Though we have examined thousands of images, the method scales well to millions of images, in which case the results would be much more robust. At the same time, we plan on comparing multiple locations. UGI would look very different for a less touristy destination than New York City. It will be interesting to observe differences across different cities and locations. We can also take the performance of individual hotels to the next level to

explore hotel competition.

Another future direction that needs to be explored is the comparison of these results to a similar clustering for the textual component of the reviews. In fact, we plan on conducting the same analysis for text, as well as the other information that is presented to users of these platforms to investigate what information a consumer uses from different data sources. A comparison should be made to the visual content generated by the hotels as well. In a broader scope this could be extremely important, because most methods for consumer search and information diffusion focus on a single modality, generally with few variables. We know that users look at both the text and images at the same time when making decisions, so understanding these interactions is imperative.







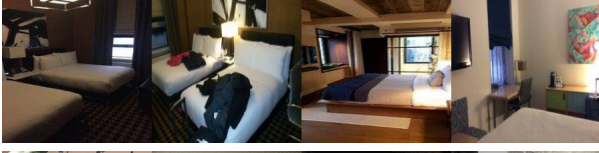
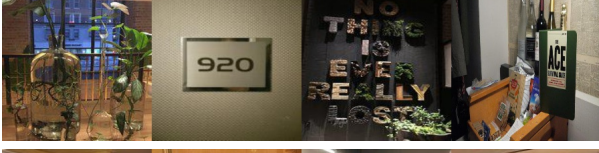
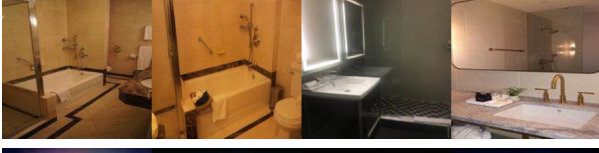

In general, we presented a useful visual analytics framework to process imagery at a large scale. The unstructured nature of imagery makes it difficult to mine opinions or consumer perceptions. We have shown how marketers can use the deep embedded clustering method to discover performance areas within UGI that are part of the consumer reviews and how to use this to investigate where they are performing well and in what areas they could improve in comparison to competition. Other studies that utilize visual analytics to solve business problems already highlighted the usefulness of these methods, but they generally need some kind of human feedback or coding to be effective. This unsupervised method is easily scalable and adaptable to larger datasets and different domains, without requiring assumptions about the distribution or underlying mechanisms.

In conclusion, visual analytics has many promising applications. Ordenes and Zhang [14] highlight several exciting directions and marketing applications such as shopping in Amazon's cashless stores (Amazon Go) or Zillow's pricing algorithms. Here visual analytics is used by Amazon to detect and automatically check out the items consumers put in their bags and by Zillow to automatically adjust prices based on detected objects in imagery. For electronic marketing specifically, research and marketers alike can use visual analytics, such as the method presented in this research, to understand consumer perceptions and opinions expressed through UGI. This is not limited to online travel or review platforms, but can also be applied to restaurants or other products and it can be used on platforms such as eBay, Amazon or Social Media. Another area that we envision to be worth exploring in this space is the automatic generation of effective visual marketing stimuli based on these new insights.

References

- [1] S. Park, Y. Yin, and B.-G. Son, "Understanding of online hotel booking process: A multiple method approach," *Journal of Vacation Marketing*, vol. 25, no. 3, pp. 334–348, 2019.
- [2] I. Chakraborty, M. Kim, and K. Sudhir, "Attribute sentiment scoring with online text reviews: Accounting for language structure and attribute self-selection," *Available at SSRN 3395012*, 2019.
- [3] Y. Wang, A. Chaudhry, and A. Pazgal, "Do online reviews improve product quality? evidence from hotel reviews on travel sites.," *Evidence from Hotel Reviews on Travel Sites.(January 22, 2019)*, 2019.
- [4] Y. Li and Y. Xie, "Is a picture worth a thousand words? an empirical study of image content and social media engagement," *Journal of Marketing Research*, p. 0022243719881113, 2017.
- [5] M. Zhang and L. Luo, "Can user generated content predict restaurant survival: deep learning of yelp photos and reviews," *Available at SSRN 3108288*, 2018.
- [6] Y. Ma and Q. Li, "A weakly-supervised extractive framework for sentiment-preserving document summarization," *World Wide Web*, vol. 22, no. 4, pp. 1401–1425, 2019.
- [7] S. Zhang, D. Lee, P. V. Singh, and K. Srinivasan, "How much is an image worth? airbnb property demand estimation leveraging large scale image analytics," *Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics (May 25, 2017)*, 2017.
- [8] L. Liu, D. Dzyabura, and N. Mizik, "Visual listening in: Extracting brand image portrayed on social media," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [9] G. Overgoor, W. Rand, and W. Van Dolen, "The champion of images: Understanding the role of images in the decision-making process of online hotel bookings," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.
- [10] Y. Ma, Z. Xiang, Q. Du, and W. Fan, "Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep learning," *International Journal of Hospitality Management*, vol. 71, pp. 120–131, 2018.
- [11] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, pp. 478–487, 2016.
- [12] J. Berger, A. Humphreys, S. Ludwig, W. W. Moe, O. Netzer, and D. A. Schweidel, "Uniting the tribes: Using text for marketing insight," *Journal of Marketing*, vol. 84, no. 1, pp. 1–25, 2020.
- [13] K.-Y. Goh, C.-S. Heng, and Z. Lin, "Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content," *Information Systems Research*, vol. 24, no. 1, pp. 88–107, 2013.
- [14] F. V. Ordenes and S. Zhang, "From words to pixels: text and image mining methods for service research," *Journal of Service Management*, 2019.
- [15] C.-F. Tsai, K. Chen, Y.-H. Hu, and W.-K. Chen, "Improving text summarization of online hotel reviews with review helpfulness and sentiment," *Tourism Management*, vol. 80, p. 104122, 2020.
- [16] K. Berezina, A. Bilgihan, C. Cobanoglu, and F. Okumus, "Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews," *Journal of Hospitality Marketing & Management*, vol. 25, no. 1, pp. 1–24, 2016.
- [17] Y.-H. Hu, Y.-L. Chen, and H.-L. Chou, "Opinion mining from online hotel reviews—a text summarization approach," *Information Processing & Management*, vol. 53, no. 2, pp. 436–449, 2017.
- [18] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment," *Expert Systems with Applications*, vol. 109, pp. 66–85, 2018.
- [19] C. Morin, "Neuromarketing: the new science of consumer behavior," *Society*, vol. 48, no. 2, pp. 131–135, 2011.
- [20] S. Zhang, N. Mehta, P. V. Singh, and K. Srinivasan, "Can lower-quality images lead to greater demand on airbnb?," tech. rep., Working Paper, Carnegie Mellon University, 2019.
- [21] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pp. 281–297, Oakland, CA, USA, 1967.
- [22] A. Ahani, M. Nilashi, O. Ibrahim, L. Sanzogni, and S. Weaven, "Market segmentation and travel choice prediction in spa hotels through tripadvisors online reviews," *International Journal of Hospitality Management*, vol. 80, pp. 52–77, 2019.
- [23] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [25] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in neural information processing systems*, 2003.
- [26] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," in *The Journal of Machine Learning Research*, 2008.
- [27] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *International Joint Conference on Artificial Intelligence*, 2017.
- [28] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in *International conference on neural information processing*, 2017.
- [29] G. Kalliatakis, "Keras-vgg16-places365." <https://github.com/GKalliatakis/Keras-VGG16-places365>, 2017.

Table 1. Overall UGI Clustering. 10 Clusters, with 4 sample images, mean, standard deviation and number of samples per cluster.

Cluster	Label	Samples	Mean	SD	#UGI
1	Zoom		3.93	1.25	450
2	Bar/Lobby		4.54	0.69	959
3	Seating Areas		4.54	0.77	793
4	Views		4.61	0.63	847
5	Hotel Front		4.53	0.69	841
6	Food/Drinks		4.57	0.71	737
7	Rooms		4.49	0.74	1700
8	Style Details		4.46	0.79	838
9	Bathrooms		4.36	0.86	1100
10	Empire State		4.68	0.57	768

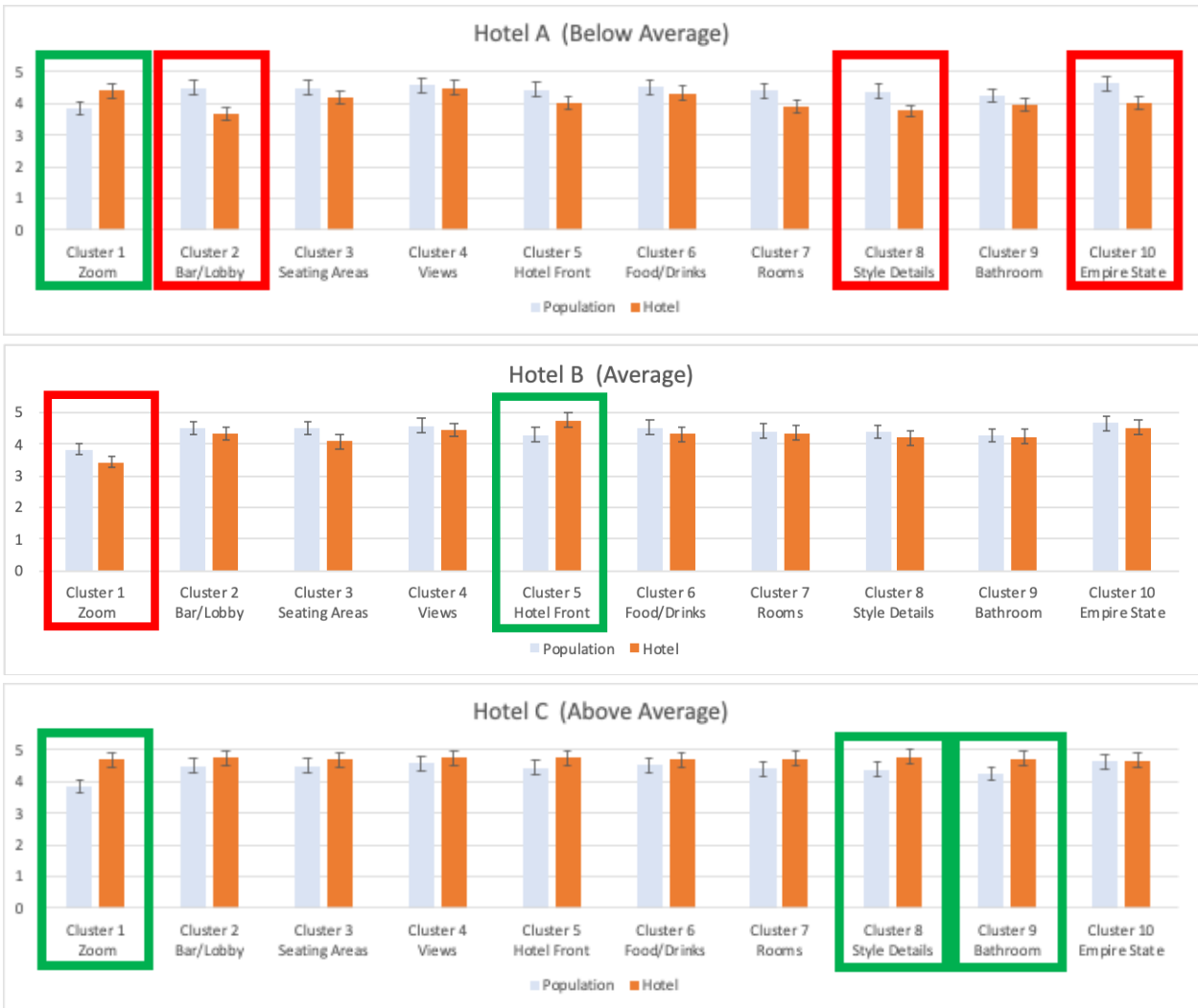


Figure 2. Average rating per cluster for 3 hotels. The selection includes a hotel with a below average rating (top), average rating (middle) and above average rating (bottom). The shaded blue bars represent the population average rating of the cluster and orange the average rating of the cluster for the hotel. The green (red) boxes indicate statistically significant positive (negative) differences from the population average.