Text vs. Image: An application of unsupervised multi-modal machine learning to online reviews

Pragna Bollam
NC State University
pbollam@ncsu.edu

Rohan Mestri
NC State University
rsmestri@ncsu.edu

Gijs Overgoor
Rochester Institute of Technology
govergoor@saunders.rit.edu

William Rand
NC State University
wmrand@ncsu.edu

Abstract

Online user-generated reviews provide a unique view into consumer perceptions of a business. Extant research has demonstrated that text mining provides insight from textual reviews. More recently, we haven seen the adoption of image mining techniques to analyze visual content as well. With data comprising of user-generated imagery (UGI) and textual reviews, we propose to perform a combination of text- and image mining techniques to extract relevant attributes from both modalities. The analysis allows for a comparison between textual and visual content in online reviews. For the UGI analysis, we use a Deep Embedded Clustering model and for the User Generated Text Analysis we use a TF-IDF based mechanism to obtain attributes and polarities. The overall goal is to extract maximum information from text and images and compare the insights we gather from both. We analyze if any modality is self-sufficient or better than the other and also if both modalities combine to give similar or contrasting insights.

1. Introduction

Before traveling, people spend hours or even days carefully selecting their choices for hotels and restaurants. With the advent of social media, online travel infomediaries like TripAdvisor and Trivago have added features on their websites that allow customers to compare prices among different hotels and check ratings [1]. As tourism increases around the globe, online websites dedicated to rating hotels have been gaining immense popularity [2]. The search costs of a customer are greatly reduced by these websites and market transparency is increased [1]. TripAdvisor is one such website that became the most popular online community about travel with more than 800 million consumer reviews and 160 million UGI. [1] TripAdvisor has added features that allow a customer to rate a hotel, write a review about it, and provide pictures. It allows users to share their experiences with comments, reviews, pictures, ratings on a hotel, destination, or any tourist attraction. These reviews have been shown to be a valuable source of information [3, 4, 5, 6, 7]

Future consumers resolve uncertainty about the quality of a hotel and its facilities by reading the online reviews given by customers on a website [1]. In addition, expectations of customers usually do not conform with their experience of a hotel because of varied reasons [2]. Consumers' decisions are impacted by online reviews, because it provides relevant insight into previous experiences of other consumers that go beyond the information provided by establishments themselves [8].

Customer satisfaction is vital for the reputation of a hotel [3]. Generally, a higher rating of a hotel implies a high reputation and customer satisfaction as these reviews provide comparative insights about customer satisfaction. A negative review can also impact the reputation of a hotel and decrease its public flow. There is a need for businesses to analyze reviews and mine the opinions of consumers, such that they can better understand how they succeed or, sometimes more importantly, fail to satisfy their patrons. The information and opinions expressed in online reviews provide valuable insights into the minds of consumers, from the perspectives of the review producers (i.e., the consumer describing the experience) and the reviewer receivers (i.e., future consumers and businesses impacted by the reviews of others) [4].

Online travel comparison websites aim to highlight the most relevant and helpful information to consumers to ease decision-making [9]. The online reviews usually include a numeric rating of a hotel, text reviews about customer experiences, user-generated images containing pictures of the hotel and other attractions, and a review with both text and imagery. In this paper, we focus on the information within these different modalities with the purpose of understanding the similarities and

[1] https://ir.tripadvisor.com/static-files /6d4c71fd-3310-48c4-b4c5-d5ec04e69d5d

URI: https://hdl.handle.net/10125/79892
978-0-9981331-5-7

differences between them and what purpose they serve from the review producer's perspective.

A recent stream of research has focused on investigating the textual content in online user-generated content such as online reviews. These studies range from opinion summarization [9, 10], sentiment analysis [11], extracting business-specific information [1, 12], or a combination such as fine-grained attribute sentiment analysis [7]. These methods are helpful in easing online decision-making by summarizing the most useful information to consumers [9]. In addition, the methods are necessary for firms to recover firm-quality information from reviews [12], or even infer quality information from what consumer do not discuss in their reviews [7].

Most of the research focused on textual reviews as 90% of the reviews on any website contain textual reviews [9], however, more frequently people are adding images along with text to share their experiences [13]. The UGI added to reviews can also be analyzed to mine consumer opinions and perceptions. The UGI, similar to their textual counterparts, hold valuable information about customer experience and hotel performance along different quality dimensions [14, 5].

Text and images hold a lot of information. The combination of both information sources has shown to be highly effective for prediction [14] or in the impact they have during online consumer decision-making [15]. For this reason, the main focus of this paper is the comparison of the textual and visual components of online reviews. The ultimate goal is to examine their differences in terms of purpose for the review producers and impact on the review receivers through an integrated methodology that combines information from both modalities.

2. Method

The goal of the paper is to extract the most talked about topics or ”attributes” from both modalities. In order to extract the most talked about topics from images, we use an Image-based clustering approach. Image-based Clustering to derive a cluster distribution is heavily discussed in [5] which relies on a combination of Transfer Learning and Deep Embedded Clustering [16] and we will modify their setup to reflect hotel-specific findings.

In terms of unsupervised text keyword extraction, several Topic Modelling Methods have been explored for document-based topic modelling [17] to short text clustering [18]. Hadifar et al [19] have also used the DEC models for text, by converting the text into SIF embeddings.

2.1. Textual Reviews

For obtaining the most important attributes from a set of reviews, we can perform a widely used method known as Term Frequency - Inverse Document Frequency (TF-IDF). TF-IDF combines two terms - Term Frequency (TF) gives the count of an attribute in a document, while Inverse Document Frequency (IDF) gives the commonality of a word in the entire document set [20].

The TF-IDF score for each attribute a in a Document D is given by

$$TFIDF_{a,D} = TF_{a,D}\ log\tfrac{N}{DF_a}$$

where $TF_{a,D}$ is the Term frequency of the attribute a in document D and DF_a is the number of documents containing the attribute a.

We start by concatenating each hotel's reviews to represent a single document. For each review that we collect, we have to preprocess the text by removing punctuations, numbers, special characters, stopwords and further lemmatize the words. We are only interested in collecting the most talked about ”attributes” for each document. These attributes are effectively the noun phrases in each sentence of the textual reviews and hence we only retain these noun phrases in the document. Once we have these attributes, we can determine the sentiment attached with them.

Attributes can also be collocations and not just a single word. Collocations are a set of words which are likely to be juxtaposed together. For example, in our case 'New York' is a set of two words which are likely to be found together as we are focusing on hotels in New York particularly. We limit ourselves to bigram collocations. We find the collocations considering all the documents as a single document and select n best collocations over the document. In case of New York, these are 'New York', 'Central Park', 'Times Square', etc. In each document, these collocations are found in an automated way and concatenated in the document.

Now we perform TF-IDF across these documents or hotels. TF-IDF highlights those attributes which are frequently mentioned about each hotel but different from the things mentioned about other hotels. We only select the top K attributes per hotel ranked by their TF-IDF scores.

Once we have these attributes, the next step would be to identify the sentiments associated with these attributes for a particular hotel. In this way, businesses can try to assess both the attributes and the polarity associated with these attributes. We try to find out the mean attribute rating and mean valence for each attribute. We try to get the mean of all the reviews in which the word is mentioned. Furthermore, the valence score (or the sentiment score) can be found out by running the rule-based Valence Aware Dictionary and Sentiment Reasoner (Vader) [21] on each sentence that the attribute is found in. We can thus find the mean rating and mean valence for each attribute in a hotel.

2.2. Image based reviews

We analyze the image-based reviews using the Deep Embedded Clustering (DEC) model. The DEC model uses a stacked denoising autoencoder that recreates the input data while reducing dimension by learning input data distribution and preserves information. This data is then passed through a decoder that reconstructs the initial input. The input is a feature vector from a pre-trained Convolutional Neural Network (CNN). It is pretrained on the Places 365 dataset [22] [23]. This VGG16 model is effective in detecting 365 common places including hotel rooms, parks, and pools. This model structurally processes images and outputs a vector of size 365 representing classification probabilities related to the 365 places it is trained to recognize. We replace this final classification layer by a sigmoid activation layer which output we can use for the DEC model. There is a three step procedure to train the DEC setup:

- We first obtain the 365 length vectors from the pretrained CNN. These features are then fed into the encoder of the DEC model, such that it is forced to replicate this vector. This is known as the pretraining phase of the autoencoder.

- The output of the encoder (”bottleneck”) is known as the latent space. Initial cluster center estimates of the data distribution are obtained by performing the k-means algorithm [24] over the latent space datapoints. The clustering layer is initialised with these cluster centres.

- After the pretraining phase and the initialization of the clustering layer, we fine-tune the model. A t-distribution based similarity is checked from each data point in the latent space to the cluster center estimates in the clustering layer. Considering q_{ij} as the t-distribution score or the ”membership” of datapoint i in cluster j

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

  where, z_i is the ith latent space datapoint, μ_j is the cluster center of jth cluster, and df is the degree of freedom of the t-distribution which is set to 1.

  The t-distribution based scores q_{ij} are trained against the t-distribution scores passed though a target distribution p_{ij}. This target distribution ”sharpens” the membership of a datapoint into its most probable cluster.

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

  A KL divergence loss metric is used for this ’self-training’. We also incorporate the reconstruction loss (the loss between the input of the encoder and the output of the decoder) in dual loss function [25], which fine-tunes the autoencoder to give out distinct cluster distributions.

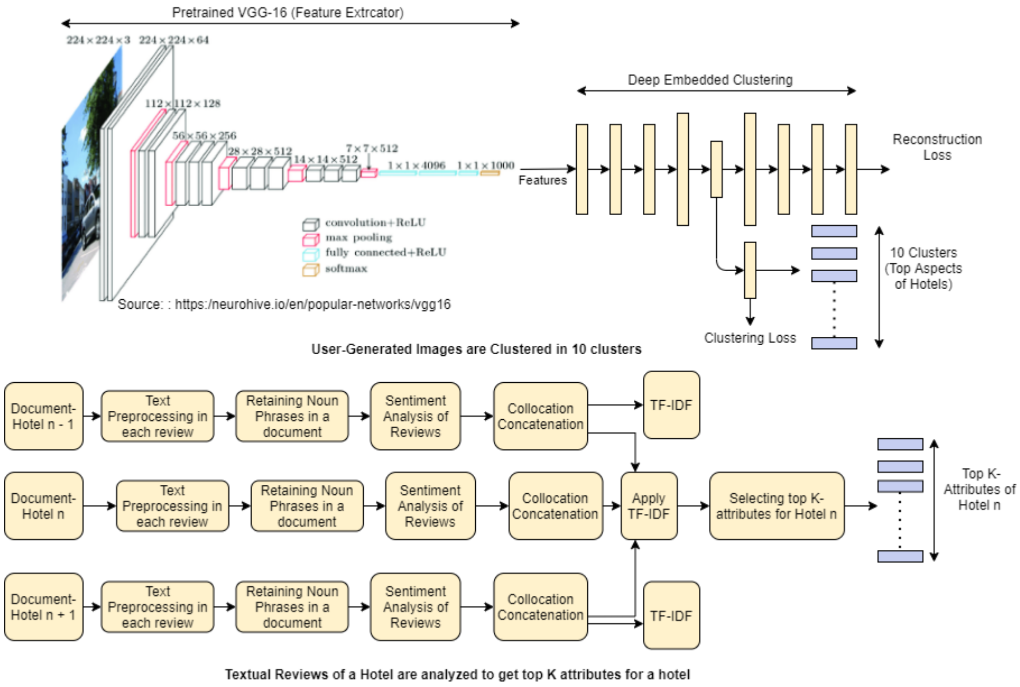
Once we have the cluster distributions, we can calculate the the mean rating of a cluster by computing the mean of all ratings of all datapoints that belong in this particular cluster [5].

3. 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 intelligence.

3.1. Data

The data used in this analysis is collected from a popular website “TripAdvisor”, the largest travel platform which has reviews for many hotels around the globe. We considered only the hotels in New York City. In total we collected 5499 online reviews with about 9144 User-generated images (UGI). Each review has a numeric rating which has an average value of 4.48/5 and a standard deviation of 0.92. The reviews on the website have positive, negative, and neutral reviews. The websites arrange the reviews in the order of usefulness from top to bottom. The textual part and the numerical rating of a review are mandatory, whereas the addition of UGI is optional. Most of the reviews found on TripAdvisor are positive reviews and about 75% of the reviews have an attached image. Of these reviews with attached images, most of them have less than 5 images attached, but there are some reviews



Pretrained VGG-16 (Feature Extrcator)
224 × 224 × 3
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000
Features
convolution+ReLU
max pooling
fully connected+ReLU
softmax
Source: : https:/neurohive.io/en/popular-networks/vgg16
Deep Embedded Clustering
Reconstruction Loss
10 Clusters (Top Aspects of Hotels)
Clustering Loss
Document- Hotel n - 1
Text Preprocessing in each review
Retaining Noun Phrases in a document
Sentiment Analysis of Reviews
Collocation Concatenation
TF-IDF
Document- Hotel n
Apply TF-IDF
Selecting top K- attributes for Hotel n
Top K- Attributes of Hotel n
Document- Hotel n + 1

User-Generated Images are Clustered in 10 clusters

Textual Reviews of a Hotel are analyzed to get top K attributes for a hotel

Figure 1. The above diagram shows the different modalities - text and images being mapped to a set of attributes, through their own methodologies.

where the maximum number of attached images went approximately to 30.

3.2. Overall Textual Reviews vs Image Reviews

Textual reviews considered for analysis are taken from reviews with and without UGI. For these textual reviews, the TF-IDF based methodology is applied to extract the top attributes. Notice that we consider all reviews to originate from one single document in this case. Effectively, this negates the IDF part in TF-IDF and we are essentially computing the Term Frequencies. These keywords extracted are the aspects of hotel services that the reviewers reviewed most frequently through text. The 10 keywords extracted are room, staff, location, bed, service, night, nyc, helpful, time, and restaurant.

Similarly, for the UGI in reviews of all hotels in New York City, a Deep Embedded Clustering with transfer learning is applied to group the images into clusters of different aspects of hotels. This method of clustering is done completely unsupervised and after clustering, each image is associated with a numeric rating of the review it belongs to. The number of clusters to be formed is chosen as 10 and the algorithm applied divided these images into clusters with high intra-class similarities. With the ratings associated with each image, an aggregate rating is calculated for each cluster. The mean and standard deviation of aggregate ratings of each cluster are taken to differentiate the cluster properties. It is observed that clusters with low average rating scores (low rating on reviews) show zoomed-in pictures of bad hotel service aspects like a broken toilet seat, bad view from the room, torn bedsheets, etc. Whereas, the clusters with high average rating scores contain images involving good hotel aspects like spacious balconies, good food, etc. The 10 clusters that were given by this approach are zoomed-in images of hotel aspects, bar, and lobby, Seating areas within the hotels, views from the hotels, front or outside of the hotels, food and drinks, hotel rooms, style details, Bathroom, Empire State Building of New York as shown in table 1.

Comparing the 10 keywords extracted from textual reviews and the 10 hotel aspects clustered from UGI, it is observed that there are some aspects like 'service' and 'staff' found in text exclusively and some aspects like 'zoomed details' and 'style details' (interiors) found exclusively through images

3.3. Hotel Specific Textual Reviews vs Image Reviews

In [5], clustering of images was not done at a hotel specific but was applied over a collection of images of various NYC hotels. We will build upon this to develop two methods where hotel-specific clustering is useful: (1) We can perform the unsupervised clustering on the entire dataset (in our case the NYC dataset). The hotel-specific datapoints can be just fed into the DEC model. In this way, the model acts as a predictor and it enforces the hotel to follow the cluster distribution of the entire dataset. (2) We can train DEC models for each of the hotels. In this way, each hotel will have its own separate cluster distribution. We prefer the former approach for the following reasons: By enforcing the hotel-specific datapoints to follow the cluster distribution of the population dataset, we define a baseline to compare hotels in the same area. In addition, it is likely that we do not have sufficient data for each hotel to train a standalone model.

For UGI belonging to a particular hotel, we first feed in all images to the VGG16 model to derive features. Then, we feed these features into the DEC to obtain the most probable cluster each datapoint belongs to. Let us call the set of image attributes which are specific to the hotels as I_{hotels}. For text, we consider each hotel's review to be a single document. This is contrary to the earlier method where we considered the entire reviews in a city to be a single document. Now, the IDF part of TF-IDF works well to eliminate out those attributes which are very common across all the documents. This reveals what is unique about each of the hotels. Let us call the set of textual attributes which are specific to the hotels as T_{hotels}

For hotel A, we observe that in T_{hotels} and I_{hotels}, the most commonly discussed aspect is the room. Consumers have posted the most images about the conditions of the 'room', 'bathroom' as well the 'view'. We can see that 'rooms' have a higher rating and 'bathrooms' is less. Similarly, from the textual analysis we can conclude that attributes like 'staff' and 'bed' is put in positive terms. In Hotel B, T_{hotels} and I_{hotels} agree on the fact that 'Bar/lobby' and 'Style Details' (interior design) has the most number of images and favorable reviews. In hotel C, 'Rooms' and 'Seating Areas' have the most number of images and have moderate mean ratings, but the textual attributes tell a different story with more emphasis on the attributes like 'nyc', 'soho' (a neighborhood in NYC), 'location'.

3.4. Text Reviews vs Text Reviews which have images attached

In this section, we try to compare the textual reviews which are associated with images $T_{hotelImages}$ as opposed to all the textual reviews for that hotel T_{hotels}. The ultimate goal is to try to understand why users add images along with their textual reviews.

In the case of Hotel B, there is not much movement in the TF-IDF scores of the attributes being talked about in $T_{hotelImages}$ and T_{hotels}. However, for Hotel C, there are new attributes like 'floor' and 'space'. Also, in T_{hotels}, staff and service is talked about more and have a lower TF-IDF score in $T_{hotelImages}$. Similarly, Hotel A has a higher TF-IDF score for visual attributes like 'bed', 'night', 'day', 'area' in $T_{hotelImages}$.

Other important thing to note is the variation of the valence score in $T_{hotelImages}$ and T_{hotels}. The valence of the word 'room' generally decreases across the three hotels, indicating that dissatisfied customers might post unfavorable pictures about the conditions of the hotel rooms. The valence of the word 'bed' drastically reduces in both Hotel B and Hotel C in $T_{hotelImages}$, implying that the users may be posting images of their dissatisfied experience with their hotel beds.

We generally observe three patterns across the three hotels discussed.

- The TF-IDF score increases/decrease for some of the keywords, implying that certain attributes are talked about more when they are attached with images.
- The valence of an attribute increases. This may imply that consumers share images that positively display a particular attribute.
- The valence of an attribute decreases. This may imply that consumers share images that negatively display a particular attribute.

4. Discussion

The surge of unstructured data sources has made methods for translating textual and visual content into useful information increasingly necessary. Online reviews have shown to be an incredibly rich source of unstructured information about consumer perceptions and experiences that provide a peek into the minds of its producers and receivers [3, 4, 5, 6, 7]. In this research, we sought to improve our understanding of the different modalities that constitute an online review - text and imagery. We presented a combination of state-of-the-art text- and image mining methods and demonstrated how these tool can be used to examine what consumers discuss when reviewing hotels. The textual analysis consisted of a TF-IDF based mechanism to extract the most frequently discussed attributes and their corresponding valence. The visual analysis consisted of a DEC model to extract the most frequently portrayed attributes by the UGI in a similar fashion. The results demonstrated clear similarities and differences between what was discussed in the text versus what was portrayed in the images. These findings persist for the general analysis of all reviews as well as the hotel-specific analysis. In line with previous research, we can conclude that text- and imagery are complementary and that mining opinions and summarizing consumer perceptions is most accurate when they are examined together [15, 9].

However, despite the stark differences between text and imaged-based reviews, they do have a few things in common. People are more likely to type textual reviews, since the current platforms encourage them to do so [2], but many consumers also feel the need to include images to highlight specific attributes or aspects that can either only be defined accurately by images or that need images to be emphasized. This is evident as most of the images are around aspects like the ’view’, ’location’, ’zoomed details’, pictures of specialty ’foods & drinks’. The main differences in the textual review analysis are attributes such as ’staff’, ’service’, ’check-in/check-out’, or perception, which are attributes that are important to the experience but that can only be described through text.

The textual reviews are also dominated by the ’room’ attribute which has a high term frequency (hence dominating the TF-IDF scores in almost all hotels), which implies that a majority of the users (mainly) talk about the hotel rooms. The image-based attributes are largely dominated by images of ’rooms’, ’seating areas’, ’bathrooms’, ’interior details’ and ’views’, demonstrating the complementarity of the two modalities from the review producer’s perspective. The discussion of rooms in both modalities indicates the need to emphasize certain elements described in text through portrayal in an image and vice versa.

Even with this analysis of the relationship between text and images within online reviews, there still exist several limitations. First, the text-based attributes have a valence score attached to them, while the image-based attributes do not. This is because textual analysis has developed extensive research in sentiment analysis, and there are several models that accurately capture human sentiments. But there is only some work on visual sentiment [26], and it is mostly related to visual content on social media. Social media imagery naturally have a large variation of sentiment compared to the mostly inanimate objects we observe in the UGI of reviews. For this reason, in future work, we create an integrated approach that combines text and images to determine what is being discussed and what the specific corresponding valence is [7]. Second, further research is needed to demonstrate the mechanisms and motivations behind consumers describing different attributes through text or image, or combining them by discussing attributes in both. This is empirically challenging from both methodological and data collection perspective. The current data consists of all textual reviews and all UGI, collected separately. The challenge is to collect both modalities for each individual review, which we have not been able to do for this version of the research, because of the way it is stored on the platform. It will certainly be necessary to truly uncover the differences between the text and the images of individual reviews in a fully integrated methodological approach. Finally, the next steps include going beyond the review producers perspective and instead focus on the impact that the different modalities of online reviews have on future consumers and businesses. Once we have a better understanding of the consumers describing their experiences, we can utilize and/or adapt our framework to understand the differential impacts of the individual and combined components of an online review on the review receivers.

In conclusion, we have made important steps towards a holistic understanding of online reviews. We have uncovered interesting dynamics between text and imagery that improve our knowledge about consumers and how they share experiences. Our multimodal methodology can be utilized for widespread future marketing research and applications trying to utilize the vast amounts of unstructured data generated by firms and their consumers in online environments.

[2] TripAdvisor, and most other platform, require textual content in addition to providing a star rating.

References

[1] N. Paolo, R. Elisabetta, and E. Paolucci, “Are customers’ reviews creating value in the hospitality industry? exploring the moderating effects of market positioning,” *International Journal of Information Management*, vol. 36, pp. 1133–1143, 2016.

[2] M. Geetha, S. Sinha, and S. Pratap, “Relationship between customer sentiment and online customer ratings for hotels - an empirical analysis,” *Tourism Management*, vol. 61, pp. 43–54, 2017.

[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] 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.

[5] G. Overgoor, R. Mestri, and W. Rand, “In the eye of the reviewer: An application of unsupervised clustering to user generated imagery in online reviews,” in *Proceedings of the 54rd Hawaii International Conference on System Sciences*, 2021.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Z. Florian, M. Yufeng, and A. F. Edward, “A preliminary analysis of images in online hotel reviews,” *e-Review of Tourism Research (eRTR)*, vol. 16, no. 2/3, 2019.

[11] 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.

[12] B. Hollenbeck, “Online reputation mechanisms and the decreasing value of chain affiliation,” *Journal of Marketing Research*, vol. 55, no. 5, pp. 636–654, 2018.

[13] M. Yufeng, X. Zheng, D. Qianzhou, and F. Weiguo, “Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning,” *International Journal of Hospitality Management*, vol. 71, pp. 120–131, 2018.

[14] Y. Ma, Z. Xiang, Q. Du, and W. Fan, “Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning,” *International Journal of Hospitality Management*, vol. 71, pp. 120–131, 2018.

[15] 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.

[16] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in *International conference on machine learning*, pp. 478–487, 2016.

[17] D. Blei, A. Ng, and J. Michael, “Latent dirichlet allocation,” *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[18] J. Yin and J. Wang, “A dirichlet multinomial mixture model-based approach for short text clustering,” *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 233–242, 2014.

[19] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, “A self-training approach for short text clustering,” *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 194–199, 2019.

[20] Q. Shahzad and A. Ramsha, “Text mining: Use of tf-idf to examine the relevance of words to documents,” *International Journal of Computer Applications*, vol. 181, no. 1, 2018.

[21] C. J. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text.,” *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pp. 194–199, 2014.

[22] G. Kalliatakis, “Keras-vgg16-places365.” `https://github.com/GKalliatakis/Keras-VGG16-places365`, 2017.

[23] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[24] 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.

[25] 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.

[26] R. Rietveld, W. van Dolen, M. Mazloom, and M. Worring, “What you feel, is what you like influence of message appeals on customer engagement on instagram,” *Journal of Interactive Marketing*, vol. 49, pp. 20–53, 2020.

Table 1. Overall Clustering of UGI for NYC hotels

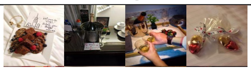
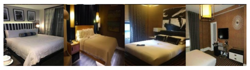
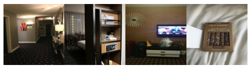
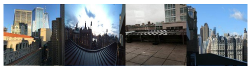
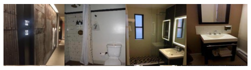
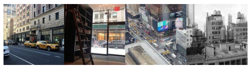
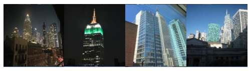
Cluster	Label	Images	Mean	Cluster	Label	Images	Mean
1	Zoomed Details		3.84	6	Food/Drinks		4.53
2	Bar/Lobby		4.5	7	Rooms		4.41
3	Seating Areas		4.51	8	Style Details		4.39
4	Views		4.58	9	Bathrooms		4.26
5	Hotel Front		4.44	10	Empire State		4.65

Table 2. Overall attributes from Textual Reviews for NYC hotels

Keyword	TFIDF	Min Rating	Max Rating	Mean Rating	Min Valence	Max Valence	Mean Valence
room	0.61	1	5	3.63	-0.67	0.95	0.24
staff	0.26	2	5	4.12	-0.71	0.98	0.54
location	0.19	1	5	3.97	-0.62	0.95	0.32
bed	0.16	1	5	3.59	-0.48	0.95	0.34
service	0.14	1	5	3.11	-0.51	0.83	0.17
night	0.14	1	5	3.53	-0.68	0.95	0.13
nyc	0.14	2	5	4.14	-0.25	0.88	0.29
helpful	0.13	2	5	4.18	-0.70	0.98	0.60
time	0.12	1	5	4.08	-0.46	0.98	0.16
restaurant	0.19	1	5	3.96	-0.44	0.92	0.23

Figure 2. Image Counts, Mean and Standard Deviation Distribution for each cluster in Hotel A

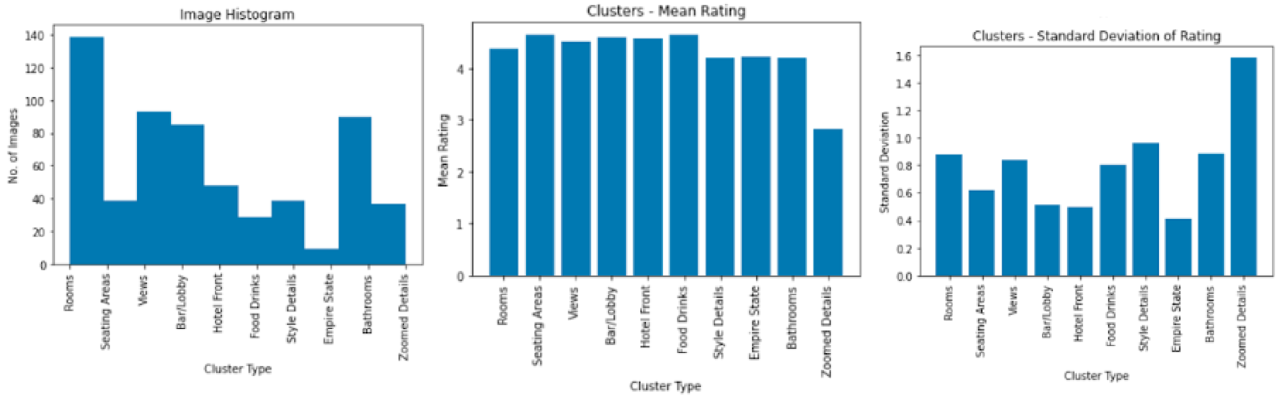


Table 3. Hotel 'A' Textual reviews only

Keyword	TFIDF	Min Rating	Max Rating	Mean Rating	Min Valence	Max Valence	Mean Valence
room	0.55	1	5	3.98	-0.93	0.99	0.28
hotel A	0.40	1	5	4.25	-0.86	0.99	0.24
staff	0.23	1	5	4.39	-0.86	0.99	0.52
location	0.18	1	5	4.36	-0.81	0.98	0.38
nylo	0.16	1	5	4.40	-0.81	0.97	0.30
bed	0.15	1	5	4.25	-0.78	0.97	0.45
restaurant	0.15	1	5	4.42	-0.82	0.97	0.25
nyc	0.14	1	5	4.43	-0.87	0.97	0.26
centralpark	0.12	1	5	4.50	-0.69	0.97	0.26
service	0.12	1	5	4.01	-0.92	0.98	0.28

Figure 3. Image Counts, Mean and Standard Deviation Distribution for each cluster in Hotel B

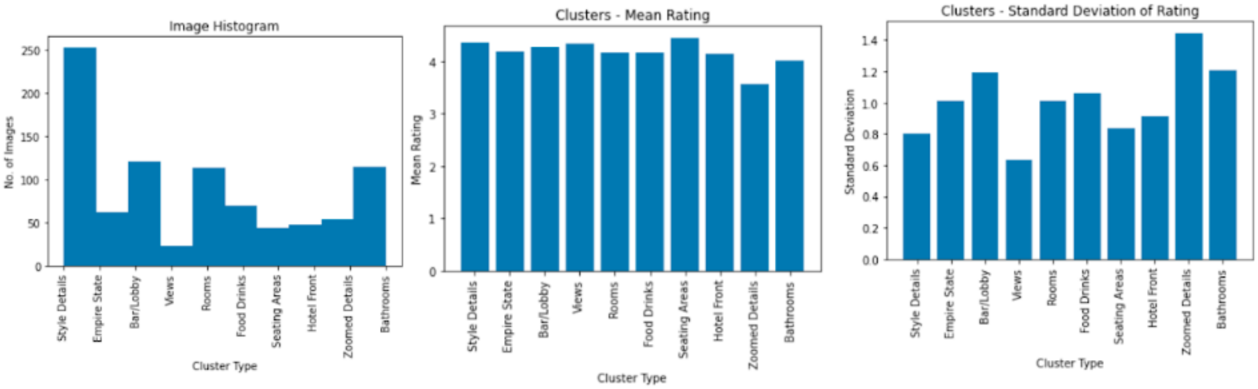


Table 4. Hotel 'B' Textual reviews only

Keyword	TFIDF	Min Rating	Max Rating	Mean Rating	Min Valence	Max Valence	Mean Valence
room	0.60	1	5	4.02	-0.82	0.98	0.27
hotel B	0.42	1	5	4.22	-0.79	0.98	0.28
lobby	0.19	1	5	4.14	-0.82	0.99	0.36
coffee	0.17	1	5	4.23	-0.81	0.99	0.37
night	0.15	1	5	3.87	-0.83	0.97	0.21
staff	0.15	1	5	4.31	-0.86	0.99	0.52
place	0.14	1	5	4.09	-0.69	0.99	0.36
nyc	0.12	1	5	4.31	-0.76	0.95	0.32
bed	0.12	1	5	4.07	-0.76	0.98	0.40
service	0.11	1	5	4.02	-0.79	0.99	0.34

Table 5. Hotel 'C' Textual reviews only

Keyword	TFIDF	Min Rating	Max Rating	Mean Rating	Min Valence	Max Valence	Mean Valence
room	0.62	1	5	3.96	-0.77	0.98	0.26
staff	0.25	1	5	4.41	-0.58	0.98	0.50
service	0.21	1	5	4.10	-0.71	0.97	0.33
hotel C	0.19	1	5	4.40	-0.78	0.98	0.28
location	0.15	1	5	4.49	-0.62	0.96	0.43
soho	0.15	1	5	4.52	-0.70	0.96	0.36
nyc	0.13	1	5	4.54	-0.69	0.94	0.28
check	0.13	1	5	4.27	-0.58	0.95	0.31
night	0.13	1	5	3.78	-0.66	0.92	0.20
time	0.12	1	5	4.19	-0.72	0.94	0.29

Figure 4. Image Counts, Mean and Standard Deviation Distribution for each cluster in Hotel C

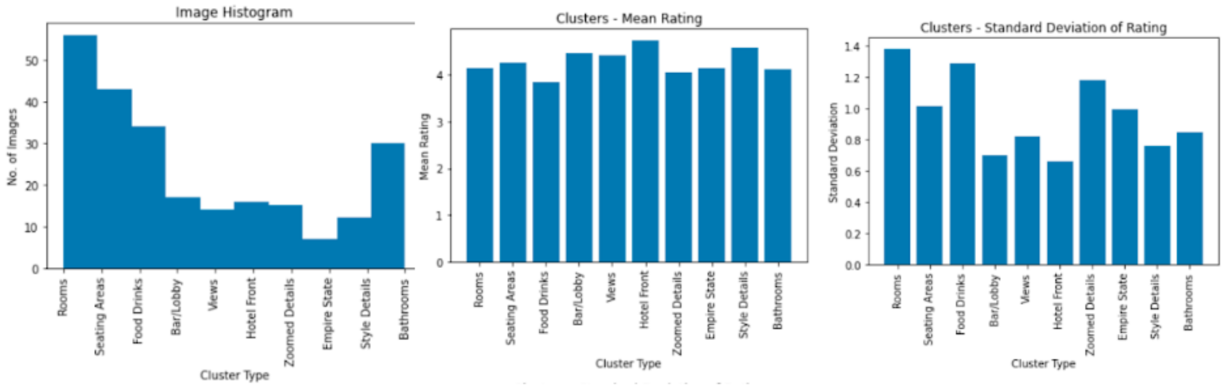


Table 6. Hotel 'A' Textual reviews with images

Keyword	TFIDF	Min Rating	Max Rating	Mean Rating	Min Valence	Max Valence	Mean Valence
room	0.59	1	5	3.91	-0.87	0.88	0.24
staff	0.21	1	5	4.54	-0.87	0.93	0.55
bed	0.19	1	5	4.24	-0.77	0.87	0.34
night	0.15	1	5	4.17	-0.78	0.94	0.28
time	0.15	1	5	3.83	-0.69	0.91	0.14
location	0.14	2	5	4.31	0	0.89	0.47
area	0.14	2	5	4.16	-0.27	0.87	0.32
nyc	0.14	1	5	4.40	-0.78	0.95	0.35
nylo	0.14	2	5	4.41	-0.77	0.89	0.27
day	0.13	1	5	4.11	-0.54	0.79	0.19

Table 7. Hotel 'B' Textual reviews with images

Keyword	TFIDF	Min Rating	Max Rating	Mean Rating	Min Valence	Max Valence	Mean Valence
room	0.67	1	5	4.08	-0.79	0.97	0.25
hotel B	0.20	1	5	4.32	-0.79	0.96	0.31
lobby	0.19	1	5	4.19	-0.64	0.97	0.28
coffee	0.18	2	5	4.34	-0.64	0.97	0.28
night	0.17	1	5	3.90	-0.69	0.94	0.19
bed	0.16	1	5	4.13	-0.75	0.97	0.39
place	0.12	1	5	4.11	-0.57	0.94	0.31
staff	0.12	1	5	4.27	-0.86	0.97	0.51
nyc	0.10	1	5	4.38	-0.48	0.95	0.31
time	0.10	1	5	4.04	-0.48	0.93	0.13

Table 8. Hotel 'C' Textual reviews with images

Keyword	TFIDF	Min Rating	Max Rating	Mean Rating	Min Valence	Max Valence	Mean Valence
room	0.61	1	5	3.78	-0.73	0.98	0.22
hotel C	0.24	2	5	4.44	-0.71	0.98	0.36
staff	0.22	2	5	4.22	-0.47	0.98	0.41
check	0.15	2	5	4.41	-0.23	0.96	0.39
soho	0.15	3	5	4.48	-0.40	0.93	0.39
location	0.13	2	5	4.35	0	0.94	0.38
service	0.12	2	5	3.88	-0.48	0.94	0.25
floor	0.11	3	5	3.83	-0.45	0.83	0.26
space	0.11	2	5	4.24	-0.32	0.88	0.47
time	0.11	2	5	4.19	-0.48	0.89	0.45