Mining Sentiment from Hotel Reviews on Resorts in the Republic of Maldives

Ian Jeffries
University of Salford, Manchester, United Kingdom

Abstract

This paper is academic in nature and explores the use of text mining to mine sentiment from hotel reviews on Resorts in the Republic of Maldives, a South Asian country located in the Indian Ocean. The following paper explores text mining on a large set of hotel reviews to find common trends in guest sentiment. Following the steps of CRISP-DM methodology, the following will discuss methods for applying correct data mining models and comparisons between R and SAS Enterprise Miner.

1. Business Understanding

1.1 Background

According to an article in Forbes magazine, [1] 90% of the data in the world was generated in the last two years alone. There are 2.5 quintillion bytes of data created each day, and approximately 90% of the world's data is held in unstructured formats. [2] Within these unstructured formats are the thoughts and sentiments of consumers in the form of reviews and text. One of the biggest challenges facing the Data Science industry today is information retrieval from these unstructured formats.

Opinion mining from hotel reviews is one of the major tasks facing the industry, as hotels attempt to find consumer opinions within massive amounts of reviews, both from social media and third-party travel websites. Why are these companies so concerned with the sentiment of their guests? According to a survey conducted by Ady and Quadri-Felitti (2015) [3], nearly 95% of travellers viewed hotel reviews before choosing a hotel and they read an average of 6-7 reviews prior to booking.

Reviews have a tremendous impact on a hotel's future bookings, and therefore text mining is required to extract consumer sentiment. By understanding certain trends in sentiment, a hotel can focus on opportunities for improvement and gather more positive reviews in the future.

1.2 Business Objectives

The following study attempts to text mine sentiment from over 20,000 reviews on hotels in Maldives – a popular tourist island located in South Asia. Sentiment will be mined using two software tools – R and SAS Enterprise Miner. Each review will be marked as containing either a Negative, Neutral, or Positive sentiment based on the positive and negative word count within each review, and trends will be identified by hotel. Time series analysis will also be used to see if sentiment varies dramatically over time.

1.3 Related Works

Since positive customer reviews have such an impact on hotel business, the demand for appropriate mining techniques has risen in recent years. As a result, there have been many studies on text mining hotels reviews. The article *Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews* from the journal of Hospitality, Marketing and Management [4] attempts to understand what aspects of the hotel's services and amenities generate positive and negative comments. This relates to the over-arching goal of this study, which identifies the overall sentiment of any given review, enabling content analysis in further studies.

The article *Opinion mining from online hotel reviews – A text summarization approach* from the journal "Information Processing & Management" [5] takes this one step further and proposes a multi-text summarization technique specifically designed for hotel reviews. That article also considers author credibility, review rating and conflicting opinions. The data in this study does not include author or review rating, and therefore those metrics cannot be analysed at this time.

2. Data Understanding

2.1 Data Description

The dataset used in this study was provided by Dr. Mo Saraee for academic purposes. Contained within the data are 8 attributes and 21,093 rows. All attributes are type "character", apart from the "date" attribute (date) and the "Total review count" attribute (integer). The names of the attributes and whether they were used in this study can be seen in Figure 1.

All reviews relate to 104 different hotels located in the Republic of Maldives, a South Asian country located in the Indian Ocean.

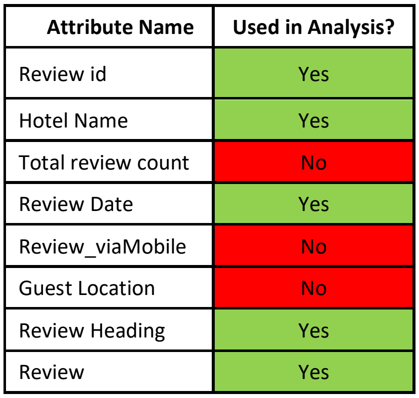
2.1 Data Exploration

A first look at the data revealed many "NA" values in the "Review_viaMobile" and "Guest Location" attributes. There appear to only be two options for "Review_viaMobile": "via mobile" or "NA". Given the scope of this study, whether the review was entered via mobile or not was deemed irrelevant and these values were subsequently ignored.

"Guest Location" appears to contain information on where the guest lives, and ranges from countries all over the world. Unfortunately, this field seems to allow manual input from the user, rather than country and city selection, and as a result many of the country values are missing or contain "NA" values. As the purpose of this study is sentiment analysis, this field was also regarded as unnecessary.

"Review Date", "Review Heading", and "Review" all contained useful information, but closer inspection revealed that there were 54 rows missing "Review" text completely – containing blanks or simply the word "More". The "Review id" and "Hotel Name" attributes seemed to be complete.

The attribute "Total review count" seemed to be complete and indicates the total review count for each hotel. Closer inspection revealed this doesn't line up, as the hotel "Robinson Club Maldives" had a total review count of 386 but an actual row count of 55. Below is the list of attributes and whether they were deemed relevant based on exploration of the data:



Attribute Name	Used in Analysis?
Review id	Yes
Hotel Name	Yes
Total review count	No
Review Date	Yes
Review_viaMobile	No
Guest Location	No
Review Heading	Yes
Review	Yes

Figure 1: Attribute Description

"Review Date" was also explored to identify the time frame of the reviews and whether any specific year had more weight. As seen in Figure 2 on the next page, the majority of the reviews occur in the 2016 – 2018 timeframe. This could indicate that there was little tracking of reviews before 2016 or that Maldives Resorts gained popularity only in recent years.

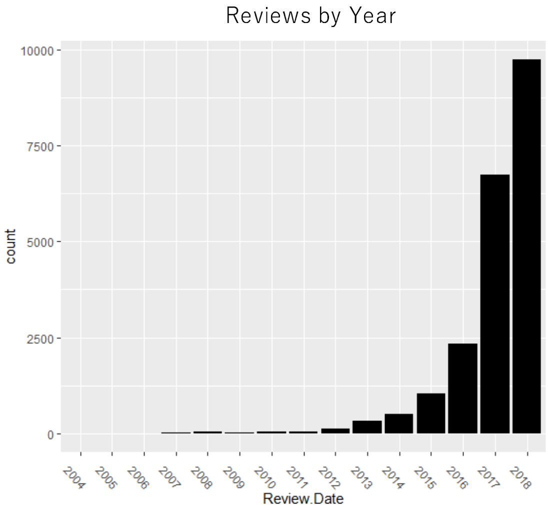


Figure 2: Hotel Reviews by Year

3. Data Preparation

3.1 Data Selection

The attributes not deemed relevant from Figure 1 were not used in analysis, which dropped the relevant attributes from 8 down to 5.

3.1 Data Cleaning

The first step in the data cleaning process was to drop all reviews that contained no information, resulting in 54 rows being removed.

Importing the data into R changed the "Review Date" field to a "character" type, and data reformatting was required to return it to a "date" format.

To successfully mine for customer sentiment, irrelevant words and characters needed to be removed from the review headers and the reviews themselves. All blank spaces were removed from the beginning and end of the reviews using the "gsub" function from R's base package. All hyperlinks, punctuations and digits were also scrubbed from the data, and all characters were converted to lower case.

The final step was to remove the commonly occurring words "Maldives", "Hotel" and "Resort", as they were not deemed helpful descriptors. All reviews are about hotels or resorts in Maldives, and therefore these words convey no new information.

4. Modelling

4.1 Model Selection

The package "wordcloud" was utilized to visualise commonly occurring words. [6] This will allow for a first pass at the overall sentiment of the reviews of Maldives. To calculate the sentiment within each review, every word within the review will be compared to positive and negative lexicons, which were retrieved from a website authored by Liu Bing. [7] By calculating the total positive or negative word count within each review, and overall sentiment percentage can be allocated. Reviews over 50% will be classified as positive.

4.1 Model Building in R

After the Maldives dataset was imported and cleaned in R, two separate corpora were created to store the header text and review body text. (Corpora are collections of documents containing natural language text) Data needed to be stored as corpora to interact with the "tm" package, [8] which will be used for text mining.

After the corpora were created, "stopwords" were removed (natural language words which have very little meaning) and the whitespace was stripped from the documents. The positive and negative lexicons were imported, and a dataframe was created to store the positive and negative word counts for each review.

A wordcloud was created to view commonly occurring words within the header text fields. (Figure 3) A first look at the overall sentiment certainly seems to be positive, with the word "paradise" occurring most frequently, followed by "amazing" and "great". Each review header was split into a list of individual words, which was compared to the positive and negative lexicons. If there

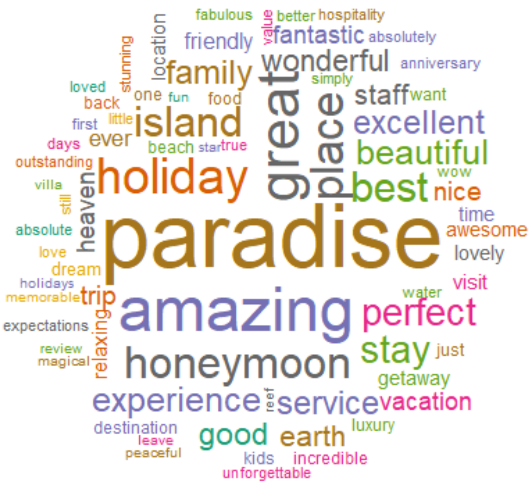


Figure 3: Review Header WordCloud

was a match in either lexicon, the positive or negative count increased for that review accordingly. A loop was created to accomplish this task for each of the 20,000+ reviews. A second wordcloud was created for the review body text. (Figure 4)

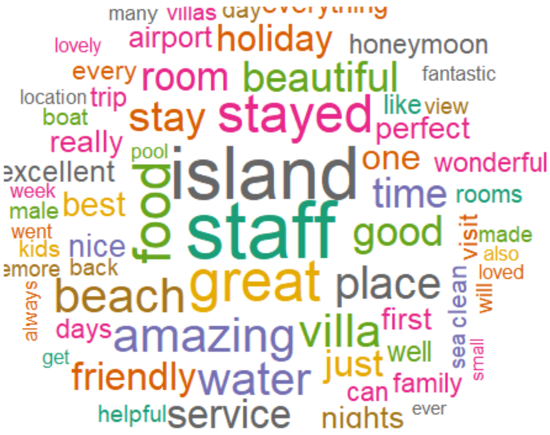


Figure 4: Review WordCloud

Two of the most frequent words were “island” and “staff”, which do not convey either positive or negative sentiment. They were removed to produce Figure 5. (“Stayed” was also deemed irrelevant and removed) Purging these words from the corpora revealed positive sentiment within the bodies of the reviews as well, with words such as “great”, “amazing” and “beautiful” occurring frequently.

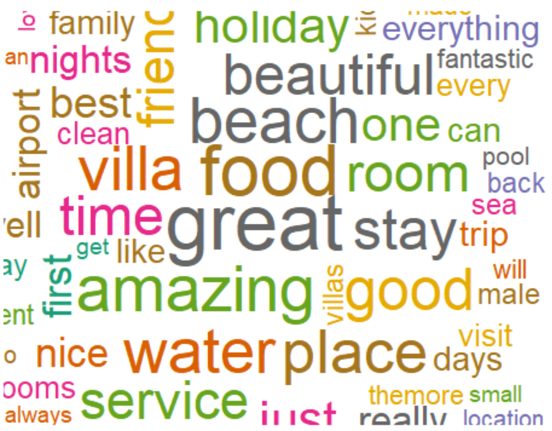


Figure 5: Revised Review WordCloud

The wordclouds give an indication of the overarching sentiment, but Figure 6 quantifies the sentiment trends between the headers and reviews. (A review was marked “positive” if the positive word count was over 50%, neutral if there was a tie or if no positive or negative words were found in the review)

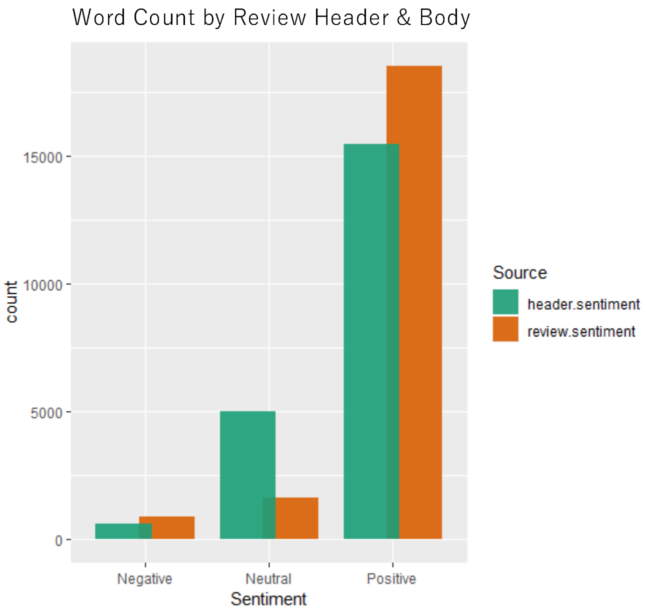


Figure 6: Header vs Review Sentiment Comparison

It is apparent that the majority of the reviews are positive, both in the header and in the review itself. To glean as much information as possible for analysis, the positive and negative word counts for the body and header within each review were combined into a single data frame, using both fields to calculate sentiment.

Within the modelling process, it was discovered that there was a large amount of duplicate reviews. The following review was repeated 29 times:

"A wonderful beautiful place to be. Ideal for 5 nights stay. Food 10/10. Service 10/10. The garden house is excellent. The on water villa are really great to try, if you feel it is expensive you can stay 1 night or 2 then move back...More"

The above review had the same Review ID and Review Date, so it's safe to assume these are duplicates. There are a lot of positive words within this review, so it's possible these duplicates are skewing the results. Figure 7 illustrates the impact of duplicate reviews.

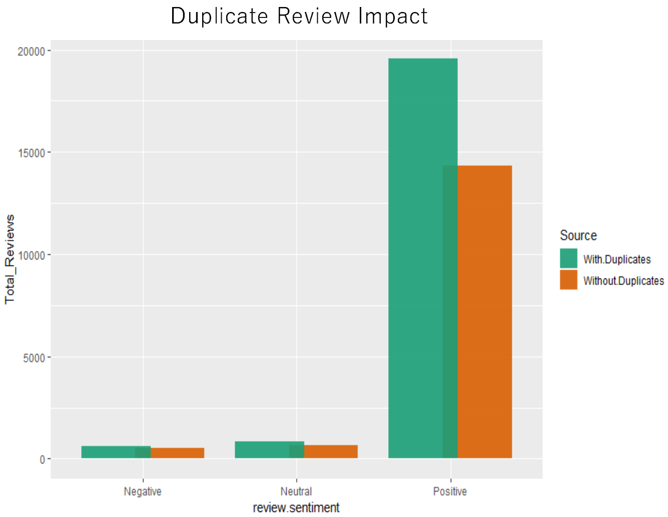
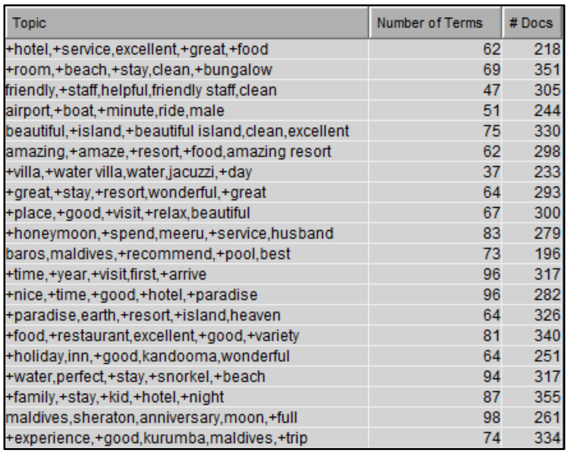


Figure 7: Duplicate Review Impact

As seen above, the duplicate reviews add a large amount of overall positive sentiment. (Almost 5,000 positive reviews) It could be possible that there are errors within the reviewing system that cause duplicate reviews, or that bots have been artificially inflating positive reviews. Even with duplicates removed, the sentiment is still overwhelmingly positive. For the integrity of this study, evaluation will only be done on unique reviews.

4.2 Model Building in SAS Miner

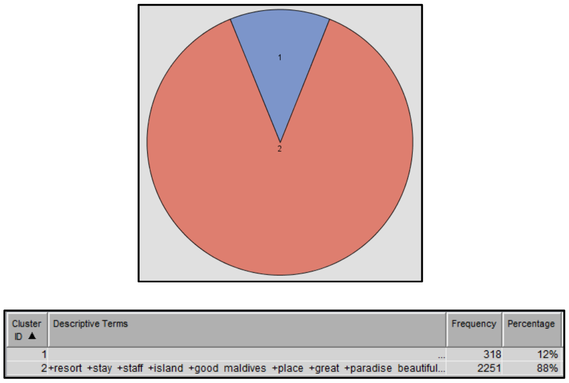
The cleaned dataset from R was imported into SAS Miner as a single column containing the header and review text combined. A text filter was added using the file "TEXT.ENGDICT", which was taken from the SAS Enterprise learning materials handed out in class. A "text topic" node was added to the diagram, and trial and error resulted in the number of "Multi-term Topics" being set to 20 in the Text Topic properties. This resulted in the following topic groups. (Figure 8)



Topic	Number of Terms	# Docs
+hotel,+service,excellent,+great,+food	62	218
+room,+beach,+stay,clean,+bungalow	69	351
friendly,+staff,helpful,friendly staff,clean	47	305
airport,+boat,+minute,ride,male	51	244
beautiful,+island,+beautiful island,clean,excellent	75	330
amazing,+amaze,+resort,+food,amazing resort	62	298
+villa,+water villa,water,jacuzzi,+day	37	233
+great,+stay,+resort,wonderful,+great	64	293
+place,+good,+visit,+relax,beautiful	67	300
+honeymoon,+spend,meeru,+service,husband	83	279
baros,maldives,+recommend,+pool,best	73	196
+time,+year,+visit,first,+arrive	96	317
+nice,+time,+good,+hotel,+paradise	96	282
+paradise,earth,+resort,+island,heaven	64	326
+food,+restaurant,excellent,+good,+variety	81	340
+holiday,inn,+good,kandooma,wonderful	64	251
+water,perfect,+stay,+snorkel,+beach	94	317
+family,+stay,+kid,+hotel,+night	87	355
maldives,sheraton,anniversary,moon,+full	98	261
+experience,+good,kurumba,maldives,+trip	74	334

Figure 8: SAS Miner Text Topics

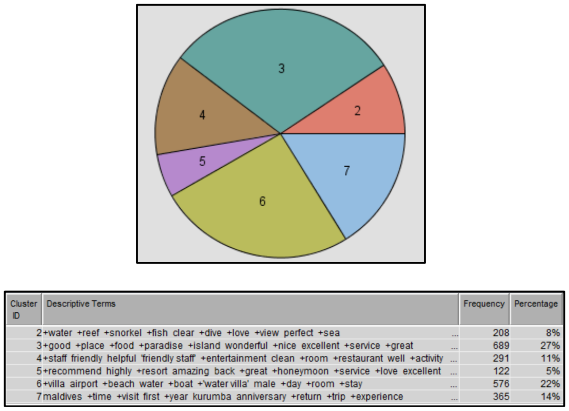
After topic groups were created, a "Text Cluster" node was added to sort commonly occurring words into groups. A first attempt at clustering set the number of clusters to 5 and the number of descriptive terms within each cluster to 10. The result was two clusters, only one of which was useful. (Figure 9)



Cluster ID ▲	Descriptive Terms	Frequency	Percentage
1	...	318	12%
2	+resort +stay +staff +island +good maldives +place +great +paradise beautiful...	2251	88%

Figure 9: SAS Miner Text Cluster

A second pass resulted in 6 useful clusters. (The first cluster was ignored, as it was pulling in blank values)



Cluster ID	Descriptive Terms		Frequency	Percentage
2	+water +reef +snorkel +fish clear +dive +love +view perfect +sea	...	208	8%
3	+good +place +food +paradise +island wonderful +nice excellent +service +great	...	689	27%
4	+staff friendly helpful 'friendly staff' +entertainment clean +room +restaurant well +activity	...	291	11%
5	+recommend highly +resort amazing back +great +honeymoon +service +love excellent	...	122	5%
6	+villa airport +beach water +boat +'water villa' male +day +room +stay	...	576	22%
7	maldives +time +visit first +year kurumba anniversary +return +trip +experience	...	365	14%

Figure 10: SAS Miner Text Cluster

As seen in Figure 10, many clusters contain positive sentiment, with the top cluster (3) containing positive words such as “good”, “paradise”, “wonderful” and “excellent”. The full process flow for SAS Miner can be seen in Figure 11.

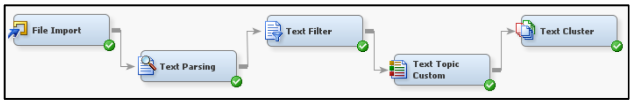


Figure 11: SAS Miner Process Flow

4.3 Model Assessment

Modelling in both R and SAS Miner revealed positive sentiment within the Maldives Hotel Reviews. R has the ability to quantify the sentiment, assigning either a positive, neutral or negative label to each review. SAS Miner allows the clustering of commonly occurring terms and creates visualisations of terms that occur frequently together. The strengths of both tools allow for a comprehensive look at the overall sentiment of the review dataset.

5. Evaluation

5.1 Evaluate Results

Text mining in both SAS Miner and R reveal overwhelmingly positive sentiment within the Maldives hotel reviews. As seen in Figure 6, a clear majority of reviews contained “positive sentiment” based on their positive word count, even with the duplicate reviews removed. Many of the top clusters from Figure 10 included positive sentiment. Using SAS Miner, concept links around specific terms can also be visualized:

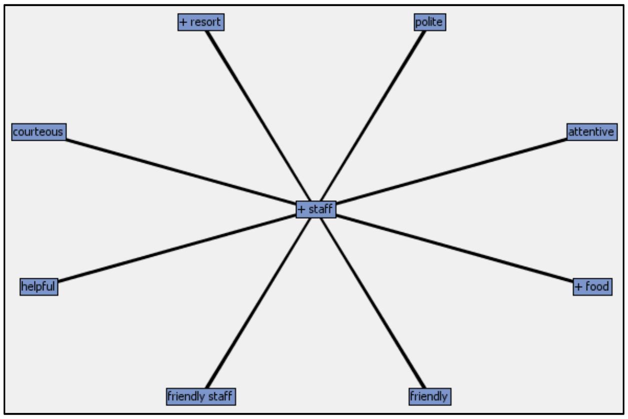


Figure 12: “Staff” Concept Links

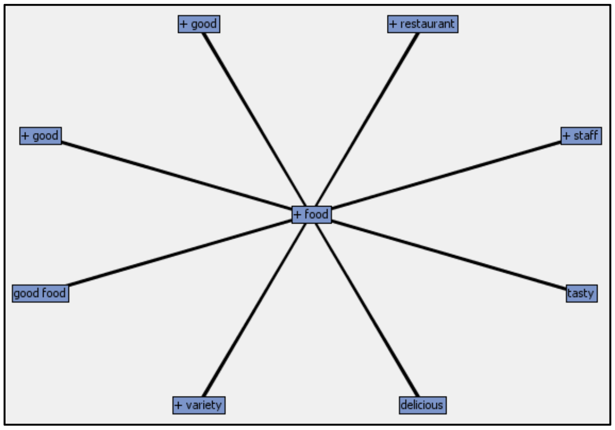


Figure 13: “Food” Concept Links

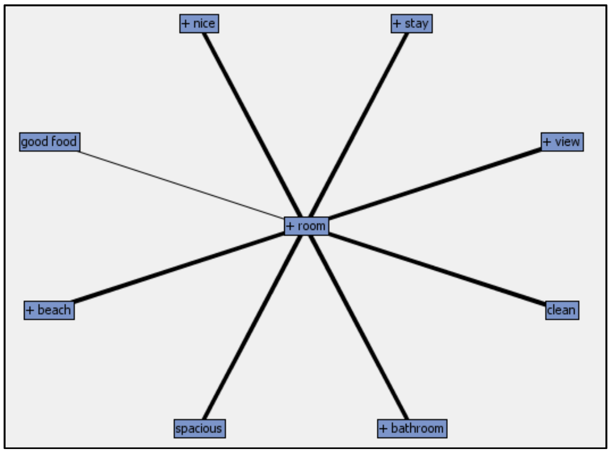


Figure 14: “Room” Concept Links

As seen in the figures above, sentiment can be mined around the most important aspects of a hotel stay: staff, food and hotel room.

According to Figure 12, common sentiment around the staff is that they're "attentive", "friendly", "helpful" and "polite". Figure 13 states that the food was "good", "tasty", and "delicious". Finally, the room was "clean", "spacious", and "nice". Presumably there was also a view.

Using time series analysis, sentiment by month can be visualized to see if it varies based on the time of year guests are travelling. As the majority of reviews are for the 2017 – 2018 timeframe, Figure 15 plots the positive and negative sentiment for all reviews by month for 2017 and 2018. As seen in Figure 15, positive sentiment does not vary greatly throughout the year. It seems to hover around 90% positive regardless of when the guest is traveling.

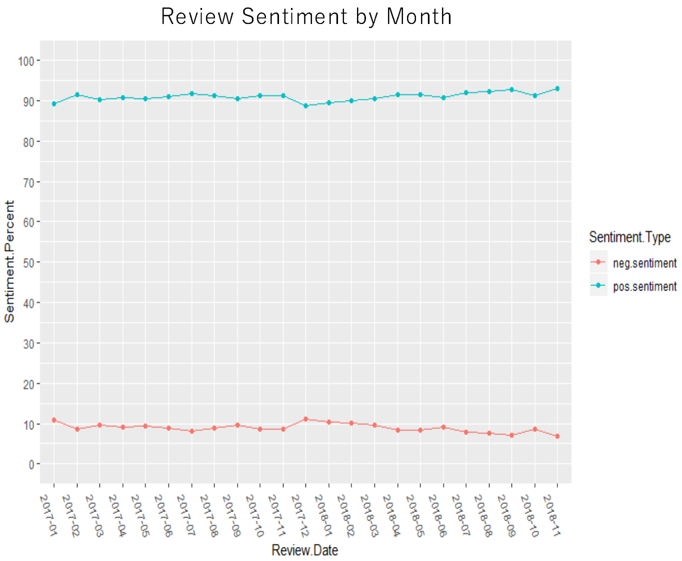


Figure 15: Sentiment by Month

The final step in evaluating guest sentiment is to look at the sentiment by hotel. Since there are 105 different hotels, this study looks at the sentiment of the top 12 hotels, determined by the largest total word count. (More reviews with a longer word count better convey total sentiment for future travellers)

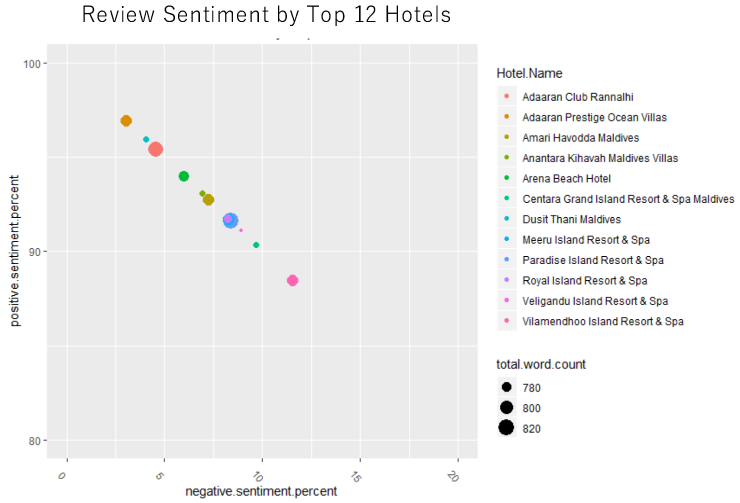


Figure 16: Sentiment by Top 12 Reviewed Hotels

As seen in Figure 16, the top hotels have an 85% positive review rate. "Alaaran Prestige Ocean Villas" had the best reviews, while "Vilamendhoo Island Resort & Spa" was on the lower end for top hotels.

In conclusion, text mining over 20,000 reviews on hotels in Maldives has revealed that it is a very nice place to visit. R calculated that 89.6% of all reviews had a positive sentiment, and SAS Miner revealed that the food, staff and room had mostly positive words associated with them.

References

[1] Marr, B. (2018, July 09). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Retrieved December 4, 2018, from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#5569cf6c60ba

[2] Saraee, M. *Text Mining: An Overview*, Lecture Notes from Applied Statistics and Data Mining

[3] Popescu, C. (2015, February 2). *Study Reveals That Travelers Prefer Summarized Review Content Over Full Text Reviews* · TrustYou. Retrieved December 4, 2018, from https://www.trustyou.com/press/study-reveals-travelers-prefer-summarized-review-content-full-text-reviews

[4] Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2015). Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. *Journal of Hospitality Marketing & Management*, 25(1), 1-24. doi:10.1080/19368623.2015.983631

[5] Hu, Y., Chen, Y., & Chou, H. (2017). Opinion mining from online hotel reviews – A text summarization approach. *Information Processing & Management*, 53(2), 436-449. doi:10.1016/j.ipm.2016.12.002

[6] Fellows, I. (n.d.). Wordcloud v2.6. Retrieved December 4, 2018, from https://www.rdocumentation.org/packages/wordcloud/versions/2.6

[7] Liu, B. (n.d.). Retrieved December 6, 2018, from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

[8] Feinerer, I. (n.d.). Tm v0.7-5. Retrieved December 4, 2018, from https://www.rdocumentation.org/packages/tm/versions/0.7-5