Data and Web Mining

Sentiment analysis of TripAdvisor Hotel Reviews

Terminal Assignment-Based Assessment

National College of Ireland

National College of Ireland
Project Submission Sheet – 2019/2020

Students Name:	Antonio Caruso, Jason Gaughan Gibbons Susana Reche Rodríguez (Group F)
Student ID:	19203608 , 19210060, 17165628
Programme:	Higher Diploma in Science in Data Analytics **Year:** 2020
Module:	Data & Web Mining
Lecturer:	John Kelly
Submission Due Date:	26/08/2020
Project Title:	
Word Count:	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examnation) if there is suspicion about the validity of their submitted work.

Signature: .....

Date: 25/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).

2. Projects should be submitted to your Programme Coordinator.

3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.

4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**

5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Table of Figures

Abstract

This paper reports a sentiment analysis performed on three 3,4 and 5 stars hotels in Dublin. The research goal was to identify any difference in consumers' perception on their stays at the hotels

The hotels reviews were scraped from the 3 selected hotels:

- The Mespil Hotel (3 star)
- The Ashling Hotel (4 stars)
- The Marker Hotel (5 stars)

By using the reviews, a ranking was built to obtain an initial understanding of the sentiment against each hotel based on four queries.

The results indicated that the Mespil was the one which differentiates the most from the others when compared against the same negative query.

After the ranking, a sentiment analysis was performed to earn more insights on the sentiment for each of the hotels. In spite of the ranking results, all the hotels seems to have pretty similar profiles. The Marker has more negative reviews related to the price. And looks like that the Mespil was the least affected by the latest COVID-19 recession as its number of reviews haven't decreased.

As a final insight, all hotels should improve the heating and acoustic insulation of the rooms.

Introduction

The project started with an initial planning on the type of information to scrape and collect from TripAdvisor (Tripadvisor, 2020). The decision veered into the hotel category in Dublin, with the analysis focused on three different hotels located in Dublin city centre respectively of 3, 4 and 5 stars. The main goal was to conduct a sentiment analysis on the TripAdvisor reviews to gain relevant insights on the consumers' experience across these three different hotels.

Below are the tasks conducted In order to obtain the final results;

1) Web scraping from the reviews section of the hotel category on the website TripAdvisor.com, with subsequent data-cleansing process applied to conduct the analysis.
2) Corpus building with subsequent ranking of the corpus, obtained by performing a total of 4 queries in the form of 1 different query per hotel and 1 common query for all the hotels for a final comparison.
3) NPL in the form of Sentiment Analysis on the corpus in order to provide insights on the hotel reviews.

Section A

Web Scraping

Decision of what to Scrape

The initial decision for the analysis was to scrape data and reviews from the hotels listed on TripAdvisor.ie, by applying the *rvest* library (Khalil and Fakir, 2017). The process began by scraping the reviews from the best seller hotels (at the time of planning) for 3 Star, 4 Star and 5 Star hotels in Dublin. The identified hotels were The Mespil Hotel, The Ashling Hotel, and The Marker Hotel. The focus was on the reviews and the date of the stay.

Web Scraping

Identify the required area within the website to scrape

They key areas were identified using a Google Chrome extension called "SelectorGadget" (SelectorGadget.com, 2013). With this tool it was possible to identify the CSS classes behind webpages which held the Page Numbers for the Reviews, the Reviews and the Date of Stay.

Get the number of Review Pages

To gather all the pages containing reviews , a function was made which takes the URL of the desired hotel and get the number of review pages. The function is using the *rvest* (Wickham [aut and [cph], 2020) and *stringr* (Wickham, 2018) libraries.

Generating the URL's for Scraping

```
> ########################################
> # Get the number of pages with reviews #
> ########################################
> getTheNumberOfReviewPages <- function(HotelURL){
+   # Read in the URL for scraping using the rvest library function "read_html"
+   webpageForPgNum <- read_html(HotelURL)
+   #webpageForPgNum
+
+   # Using the URL and the stringr library funtion "str_sub" to create a character
 vector of the information
+   # stored within the identified class on the webpage
+   pageNumber <- str_sub(webpageForPgNum %>%
+                         html_nodes(".pageNumbers") %>%
+                         html_text())
+
+   # Remove unwanted characters from the beginning of the character vector and the
n convert the remaining numbers
+   # to a numeric data type
+   pageNumber <- substring(pageNumber, 8)
+   pageNumber <- as.numeric(pageNumber)
+   print(pageNumber)
+   return(pageNumber)
+ }
> pageNumber <- getTheNumberOfReviewPages(MespilHotelURL)
[1] 240
```

Figure 1 Getting the Number of Pages for Scraping

Next in order to start generating the URL's that are needed for scraping an empty list vector is created. Then using the page numbers which were scraped previously and knowing there was 5 reviews per page, the page numbers were multiplied by 5 in order to capture the entire range of numbers needed for generating the URL's.

Once the preparation was completed, a FOR Loop was run and created a URL for each value within the sequence of numbers from 5 to the total range of page numbers. For each iteration of the loop using the glue function from the *glue* library , a URL was created and stored within the list vector. A snippet of this can be seen below [Fig.2].

```
[[5]]
https://www.tripadvisor.ie/Hotel_Review-g186605-d212567-Reviews-or1175-Mespil_Hotel
-Dublin_County_Dublin.html#REVIEWS

[[6]]
https://www.tripadvisor.ie/Hotel_Review-g186605-d212567-Reviews-or1180-Mespil_Hotel
-Dublin_County_Dublin.html#REVIEWS

[[7]]
https://www.tripadvisor.ie/Hotel_Review-g186605-d212567-Reviews-or1185-Mespil_Hotel
-Dublin_County_Dublin.html#REVIEWS

[[8]]
https://www.tripadvisor.ie/Hotel_Review-g186605-d212567-Reviews-or1190-Mespil_Hotel
-Dublin_County_Dublin.html#REVIEWS

[[9]]
https://www.tripadvisor.ie/Hotel_Review-g186605-d212567-Reviews-or1195-Mespil_Hotel
-Dublin_County_Dublin.html#REVIEWS

[[10]]
https://www.tripadvisor.ie/Hotel_Review-g186605-d212567-Reviews-or1200-Mespil_Hotel
-Dublin_County_Dublin.html#REVIEWS
```

Figure 2 Snippet of Links ready for scraping

Next in order to only have valid URL's, all the null values from the vector were removed. By using the unlist function, the list of URL's was converted to a character object. Once done, a further FOR Loop was ran to create a new list vector of the same length of character object.

Scraping of each URL to gather the Review and Date of Stay

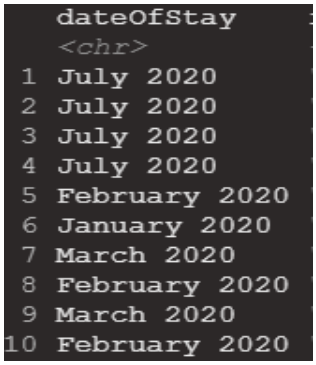
Finally, one last FOR Loop iterating through a sequence and through each URL within the character vector, using again the *rvest* library (Hadley Wickham [aut and [cph], 2020) and *stringr* library (Wickham, 2018), captured each Date of Stay and Review within each URL using the CSS classes identified with "SelectorGadget".

For each time the FOR Loop iterated, it scraped and stored the retrieved information within a data frame within the loop. Once the loop was completed and the Date of Stay and the Review were retrieved, by using the bind rows function it was then possible to take and store the retrieved information within a final data frame for further use.

	dateOfStay	review
	<chr>	<chr>
1	Date of stay: July 2020	"We stayed at the Mespil as a family of four. The staff were fantastic and nothing was too much trouble for them especi~
2	Date of stay: July 2020	"Mespil Hotel Mespil Hotel is situated right across from a lock for barges. Its on 50-60 Mespil Road, Dublin 4. This is~
3	Date of stay: July 2020	"Lovely stay in the Mespil Hotel. We found it a great location close to the City Center without being in the center of t~
4	Date of stay: July 2020	"Just back from a lovely stay at the fabulous Mespil, I had forgotten how lovely a location and setting it has but was r~
5	Date of stay: February ~	"So where do I start...The reception staff were very accommodating and super friendly with my son as soon as we arrived.~
6	Date of stay: January 2~	"Beautiful, modern, comfortable Hotel and exceptionally friendly Staff. I stayed at the Mespil Hotel for a medical Confe~
7	Date of stay: March 2020	"Everything about our stay here was perfect. Staff went out of their way to accommodate us, check in and out was effici~
8	Date of stay: February ~	"Five of us stayed for 3 nights. The reception staff could not do enough to help us with outings, restaurants etc and a~
9	Date of stay: March 2020	"Title was purely \"Clickbait\". Anyway, stayed on Friday night as was attending a wedding near Harcourt St, so was loca~
10	Date of stay: February ~	"Apologies for the long review but I just had to review the whole experience at Mespil. This isn't a chain hotel. It's ~

Figure 3 Snippet of the final Web Scraping output

Data Cleaning and Preparation for Analysis



	dateOfStay
	<chr>
1	July 2020
2	July 2020
3	July 2020
4	July 2020
5	February 2020
6	January 2020
7	March 2020
8	February 2020
9	March 2020
10	February 2020

Figure 4 Cleaned Dates of Stay

The Data frames created with the "dateOfStay" and "review" for each hotel needed to go through a degree of preparation for further analysis. This began with the "Date of Stay: " needing to be removed from the character string for the "dateOfStay" column. So again using the *stringr* (Wickham, 2018) library function str_sub the first 15 characters were removed from the string for each row in the column. Only the the month and year remained of use. For further analysis, each of the created hotels data frames were saved in individual .csv files.

Next the aim was to import each individual file and combine the data, but in doing so for clarity an additional column with the hotel name was added. The final data frame was created using the rbind function, resulting in the date of stay, reviews and the hotel name of all combined hotels. This was then exported to a .csv.

The singular .csv files for each hotel were used to build the Ranking of the Corpus, whereas the combined .csv file holding all three hotels were used for the Sentiment Analysis.

Ranking Building Process of the Corpus

Building Corpus

This section explains the process for building a Corpus, applied separately to each of the hotels. The readlines function (Study.com, 2019) was employed initially to extrapolate the reviews from the rows of each csv document and convert them into a text document. A for loop containing a list function (Fanara, 2018) was then applied to fill the text document of each hotel with all the existing reviews turned into separate documents, each containing a review. 1 pretended query per hotel was set up and incorporated as a string of text in a corpus intended as a collection of the reviews (Lantz, 2015) plus the query, resulting in 751 final documents. The tm text mining package and the Snowball package were applied (Lorna, M.A., 2018) to standardise the documents in the corpus and implement the StemDocument transformation, aiming at implementing the Porter Stemmer Algorithm (Weiss, et al., 2010) used for semantic checking and removal of irrelevant stop words, symbols and extra space.

Application of Vector Space Model for Building a Ranking of the Corpus

The Vector Space Model was employed to rank the reviews of each hotel based on the similarity of each to the three different queries applied to each hotel to obtain a small search engine (Manning, et al., 2008). The first step consisted in the creation of a TermDocumentMatrix with rows corresponding to the words and columns corresponding to the documents, including the query. The function “sqrt(tf) or log(tf)” (Silge & Robinson, 2020) was used as logarithm to build the search engine. The application of the formula ((1 + log2(tf)) $*$ log2(N/df)) implemented the weight allocation across the rows of the term document matrix, through the use of the tfidf function (Lants, 2015),which resulted in a weight applied on every row of the term document matrix to obtain the ranking. By normalising each column vector in the tfdif matrix, it was possible to obtain the dot product cos 0, where 0 was intended as the angle between two points (a) and (b), whereby the cosine decreases from the maximum value of 1.0. The matrix operation returns values of Cos0 for each document vector and the query vector.

Application of Queries and Interpretation of the Results

This section shows the application of the queries to each of the Hotels and the interpretation of the final ranking obtained by the search engines created.

Performance of a Query for The Mespil Hotel

- The query chosen for the 3 stars Mespil hotel was **"family hotel and good breakfast"** identified as potential keywords matching the hotel, and added to the corpus as following;

```
query <- "family hotel and good breakfast"

docs_Mespil <- VectorSource(c(review.list, query))

docs_Mespil$Names <- c(names(review.list), "query")

mespil_corpus <- Corpus(docs_Mespil)

inspect(mespil_corpus)
```

Figure 5 "Query and Corpus building"

- The corpus obtained was then cleaned with the tm function.

```
mespil_corpus <- tm_map(mespil_corpus, removePunctuation)
mespil_corpus <- tm_map(mespil_corpus, stemDocument)
mespil_corpus <- tm_map(mespil_corpus, removeNumbers)
mespil_corpus <- tm_map(mespil_corpus, tolower)
mespil_corpus <- tm_map(mespil_corpus, removeWords, stopwords("english"))
mespil_corpus <- tm_map(mespil_corpus, stripWhitespace)

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
mespil_corpus <- tm_map(mespil_corpus, toSpace,"â€¦")
mespil_corpus <- tm_map(mespil_corpus, toSpace,"¦")
mespil_corpus <- tm_map(mespil_corpus, toSpace,"€™")
mespil_corpus <- tm_map(mespil_corpus, toSpace,"â€“")
inspect(mespil_corpus)  #cleaning completed
```

Figure 6 "Cleaning of the text corpus"

- The TermDocument Matrix was created, indicating 14 terms appeared at least once in 751 documents.

```
<<TermDocumentMatrix (terms: 14, documents: 751)>>
Non-/sparse entries: 553/9961
Sparsity           : 95%
Maximal term length: 7
Weighting          : term frequency (tf)
Sample             :
         Docs
Terms     1 12 19 22 242 29 419 424 6 749
  back    1  1  0  2   0  1   0   2 0   2
  came    1  0  0  1   0  0   0   1 0   0
  definit 1  1  2  0   0  1   1   0 1   1
  end     1  0  0  0   0  0   0   0 1   0
  enjoy   1  0  1  1   2  0   1   0 1   1
  especi  1  0  0  0   0  0   0   0 1   0
  even    1  1  2  0   2  0   0   0 0   1
  extra   1  0  0  0   0  0   0   0 0   0
  famili  1  0  0  0   0  1   3   0 1   0
  fantast 1  1  0  0   0  1   0   2 0   0
```

Figure 7 "TermDocumentMatrix"

- The Tfidf weight from a term frequency vector and a document was applied and then run on every row of the term document matrix, whose frequency was derived from each row by counting the non-zero entries and excluding the query.

```
get.tf.idf.weights <- function(tf.vec, df){
  weight = rep(0, length(tf.vec))
  weight[tf.vec > 0] = (1 + log2(tf.vec[tf.vec > 0])) * log2(N.reviews/df)
  weight
}

get.weights.per.term.vec <- function(tfidf.row) {
  term.df <- sum(tfidf.row[1:N.reviews] > 0)
  tf.idf.vec <- get.tf.idf.weights(tfidf.row, term.df)
  return(tf.idf.vec)
}

tfidf.matrix <- t(apply(Mespil.doc.matrix.stm, c(1), FUN = get.weights.per.term.vec))
colnames(tfidf.matrix) <- colnames(Mespil.doc.matrix.stm)
```

Figure 8 "Tfidf Weight Function"

- Normalisation of each column vector.

```
angle <- seq(-pi, pi, by = pi/16)
plot(cos(angle) ~ angle, type = "b", xlab = "angle in radians", main = "Cosine similarity by angle")
```

```
tfidf.matrix <- scale(tfidf.matrix, center = FALSE, scale = sqrt(colSums(tfidf.matrix^2)))
tfidf.matrix[0:3, ]
```

Figure 9 "Normalisation of each column vector"

- Matrix Multiplication

```
query.vector <- tfidf.matrix[, (N.reviews + 1)]
tfidf.matrix <- tfidf.matrix[, 1:N.reviews]
```

Figure 10 "Matrix Multiplication"

- Values of Cos 0 for each document vector and query vector

```
doc.scores <- t(query.vector) %*% tfidf.matrix
```

Figure 11 "Values of Cos 0 for each document vector and query vector"

- Documents ranked by their cosine similarity with the query vector

```
results.df <- data.frame(doc = names(review.list), score = t(doc.scores), text = unlist(review.list))
results.df <- results.df[order(results.df$score, decreasing = TRUE), ]
```

Figure 12 "Documents ranked by their cosine similarity with the query vector"

- Result of the search engine

```
      doc     score

   text
419 doc419 0.4613302                                                                                                  Thank you to all the wonderful staff at the Mespil for looking after our f
lin so enjoyable. Special thanks to Mr. Gerard Kelly who is always so helpful and pleasant. Our family room was spacious and very comfortable. We are looking forward to a return visit a
 family.
461 doc461 0.4143245
                                      "Stayed here for 2 nights with friends & Family, just a short walk in to the city. Good value for Dublin nice breakfast and room was comfortable,
endly."
624 doc624 0.3053491 "We stayed as an extended family group in multiple rooms and our family room was great, nothing fancy but modern , decent sized and well equipped.  The breakfast wa
 and about a 15-20 minute walk to the city centre.  The staff were cheerful, friendly and nothing was too much trouble. From booking taxis, to recommendations to storing flowers - a sup
 stay."
257 doc257 0.2919168
                                      "Had a nice family room at this hotel ..., staff very pleasant , beds comfortable, easy free parking , good breakfast and reasonably priced. Easi
eturn."
277 doc277 0.2651366
                                   "Very nice stay for our family. Easy walk to restaurants. Breakfast at the hotel is a must! The staff was gracious and helpful, especially Peter. Lo
ended!"
395 doc395 0.2569651
 the rooms have been updated, great breakfast, friendly staff, and great restaurants nearby.  While we had no view, it actually made us get out and explore a little more.  Would definit
ublin."
```

Figure 13 Figure 9 "Ranking documents of the Mespil Hotel reviews"

For the query **"family hotel and good breakfast**" the output shows that the best document review that matched the query was the doc419 with the highest score of similarity of 0.46.

Performance of a Query for The Marker Hotel

- The exact same procedures was applied to the 5 stars Marker Hotel's reviews, resulting in the search engine as below for the query **"luxury hotel with spa and rooftop"**

```
       doc     score
                                                                                                    text
793   doc793 0.2947259                                                            A fantastic and luxurious stay.  Great attentive service from all staff whi
lst there and amazing facilities to enjoy. Highlights were the spa and the rooftop bar (gets busy on a sunny Friday night though!)
1362 doc1362 0.2829588                                                  The Marker is a really gorgeous hotel and we had a most luxurious stay. The location is gre
at and the rooms are very comfortable. The staff are all great and very helpful. I'd highly recommend The Marker Hotel as a treat.
1035 doc1035 0.2786855                                                        Excellent hotel and superb staff. Very relaxing environment and the food in the Bras
serie was excellent.  I would highly recommend this hotel for a luxurious weekend stay. The rooms and spa facilities are excellent.
811   doc811 0.2711698        "Would highly recommend.  The excellent 5 star service from staff really made our stay fantastic.  Spacious, clean, comfortable and luxurious hote
l.  Bedroom was perfect and the bathroom with the heated floor was amazing!  Enjoyed the breakfast and spa also. Would stay again."
1270 doc1270 0.2684890                   "Beautiful modern luxury hotel. Large comfortable room with nice view of the square and canal.  Loved the excellent AC.  Staff all f
riendly and helpful.  Good breakfast at the Brasserie.  Loved the rooftop bar, what a treat!  We would surely return if in Dublin."
409   doc409 0.2662183 "Beautiful, modern and luxurious hotel with a real 5* feel. Location ideal. Worth every penny. You will not be disappointed staying here if you enjoy luxu
ry.  The bar menu is fantastic, cocktails are to die for and the roof top bar is amazing in summer!  Also has a pretty great spa! "
```

Figure 14 "Ranking documents of the Marker Hotel reviews"

The Doc793 is the one that ranked highest with 0.29. There is a smaller gap score for the Marker Reviews compared to the Mespil, potentially because of some specific keywords like spa and rooftop which would be more specific in a review.

Performance of a Query for The Ashling Hotel

As per the previous hotels, the exact same procedures were applied to the 4 Stars Ashling Hotel's reviews, resulting in the search engine as below for the query **"clean comfortable good quality hotel"**

```
                                                text
1126 doc1126 0.2685027
                                                                                                                                                "We stayed for two nights, big room, massive bathroo
m. A nice hotel, clean and everything you need. Breakfast was good, large choice But I felt that the ingredients for the cooked breakfast could have been of better quality. This
 would have made a good breakfast, great. We would stay again."
638   doc638 0.2658083
                                                                                                                                                "Located near to the tram stop, which made for easy acc
ess into Dublin city centre.  Very clean and with kind and helpful staff.   We had a large room with all the amenities. At breakfast we were offered a wide range of quality food
 and the service was very good. We would recommend. Thank you"
1648 doc1648 0.2572297

                                We had a great room on our quality time weekend.  Bus and tram services on the doorstep. Maps provided for the city. Our room was spacious and
 clean and would recommend to anyone visiting this great city
2124 doc2124 0.2484817
                                                                                                                                                                                "Very comfortable and co
nvineant right next to the train station and, tram. Rooms are very up to date modern and clean with all the amenities. Breakfast is excellent with a lot of variety and good qual
ity. Our first choice to stay whenever we are visiting Dublin"
595   doc595 0.2424943 "We stayed two nights at the Ashling hotel. Good location near tram/train station, museum, good pubs and Guinness brewery but away from the noise and crow
d of Temple bar. The front staff was very friendly, efficient and helpful. The room was clean and spacious with a coffee maker, room safe and a tv with many stations. Some traff
ic noise in the morning. The bed was comfortable and the WiFi was good. Nice clean bathroom with a good shower. The breakfast buffet had a good selection and the food was of goo
d quality. There is also a bar In the hotel with a food menu."
1612 doc1612 0.2305689

                                  A great hotel from a very pleasant check in to a a great variety and quality breakfast. Right beside the train station too and close to the
 city centre on the Liffey. Lovely surprise after long flights.
```

Figure 15 "Ranking Documents of the Ashling Hotel Reviews"

The Doc 1126 ranked highest with 0.26 for the above query. Although the words are less specific than the previous query, the higher number of reviews could have determined a lower score in the top reviews than the Marker.

Performance of a common query for the three hotels

The Mespil Hotel	The Ashling Hotel	The Marker Hotel

Figure 16 "Comparison of Hotels Reviews Ranking on a negative query"

Since all the three positive queries matched reviews with relevant scores, a fourth negative query **"bad and uncomfortable experience"** was run on the three hotel corpora to identify their ranking and obtain insights. The output shows how the Marker and the Ashling ranked 0, whereas the Mespil presented a relevant score for negative reviews matching the negative query. Finally, before conducting the sentiment analysis, it could be assumed that reviews would be all extremely positive for the Ashling and the The Marker, whereas the Mespil might have offered some negative experience to its customers complaining on TripAdvisor.

Section B

Sentiment Analysis

Sentiment analysis was chosen over topic modelling to understand the customers' feelings about the three hotels. By using natural language processing (NPL) meaningful information was extracted and then analysed.

The lexicon dictionaries was the method used and the chosen R package "Tidytext". This type of method considers that the sentiment in the document is the sum of the sentiment across different words (Devika M D, 2016).

Similar with the ranking cleaning, the text needed to be free of punctuation, digits, stop words, words of length 1 and white spaces. The words needed to be all in lower case and also reduced to their root. Once the data was clean and in the right format (each word in one row), the data frame containing the words could be joined with the different lexicons, to understand the assigned sentiment for each of them.

The first lexicon used for the sentiment analysis was the "AFINN" [Fig.17][page.37, Appendix B].

	hotelName	date	word	value
	<chr>	<yearmon>	<chr>	<dbl>
1	The Marker Hotel	Jul 2020	impressed	3
2	The Marker Hotel	Jul 2020	reassuring	2
3	The Marker Hotel	Jul 2020	welcomed	2
4	The Marker Hotel	Jul 2020	helpful	2
5	The Marker Hotel	Jul 2020	fantastic	4
6	The Marker Hotel	Jul 2020	friendly	2
7	The Marker Hotel	Jul 2020	stunning	4
8	The Marker Hotel	Jul 2020	best	3
9	The Marker Hotel	Jul 2020	uncertain	-1
10	The Marker Hotel	Jul 2020	recommend	2

… with 33,899 more rows

Figure 17: Reviews data frame joined with AFINN Lexicon

As per [Fig.18] the Marker Hotel is the one with the highest mean, but by analysing the median [Fig.19] all of them seems similar.

hotelName	`mean(value)`
<chr>	<dbl>
The Ashling Hotel	1.91
The Marker Hotel	2.24
The Mespil Hotel	2.00

Figure 18: Mean Overall Sentiment Analysis AFIN

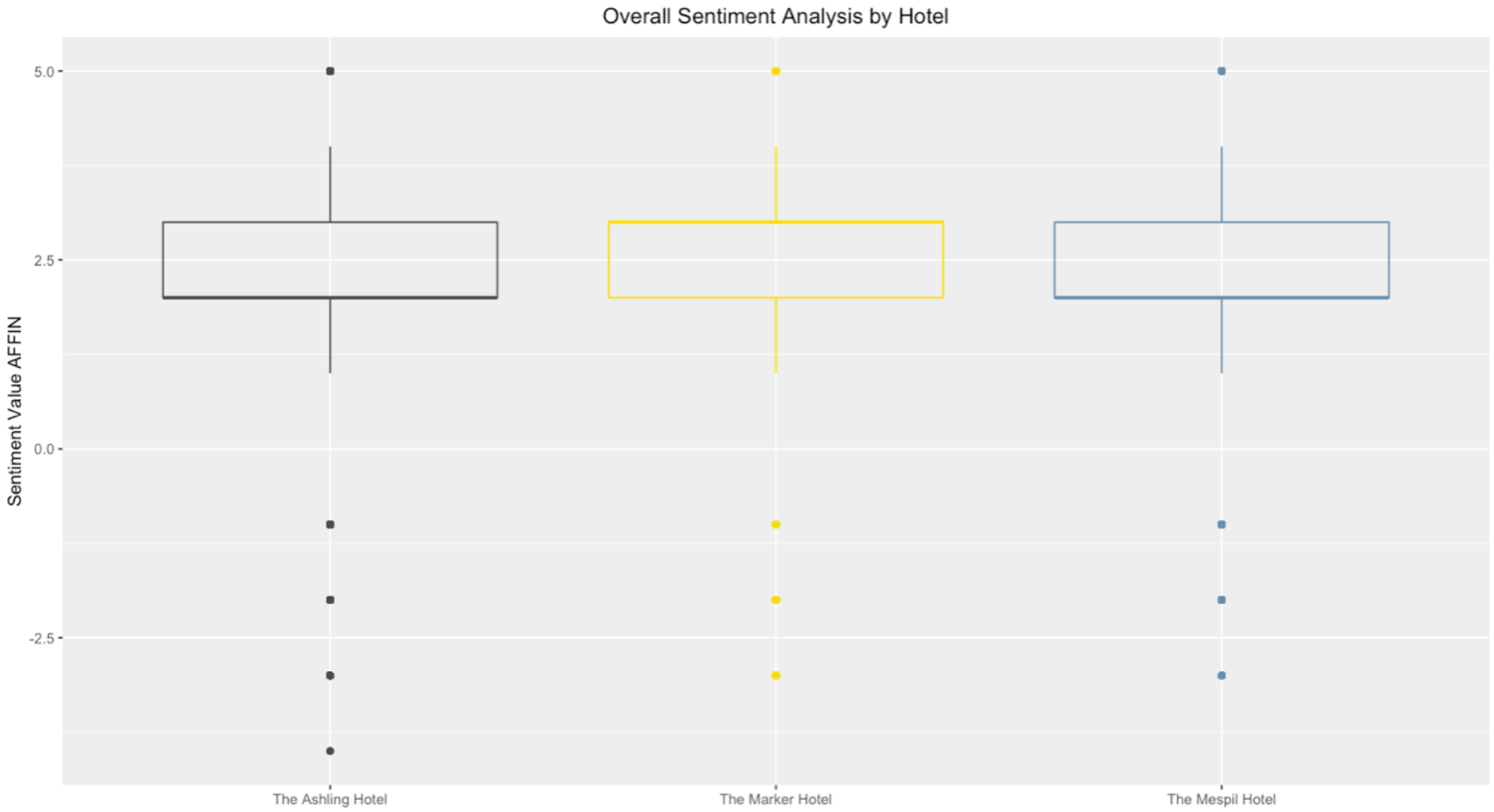
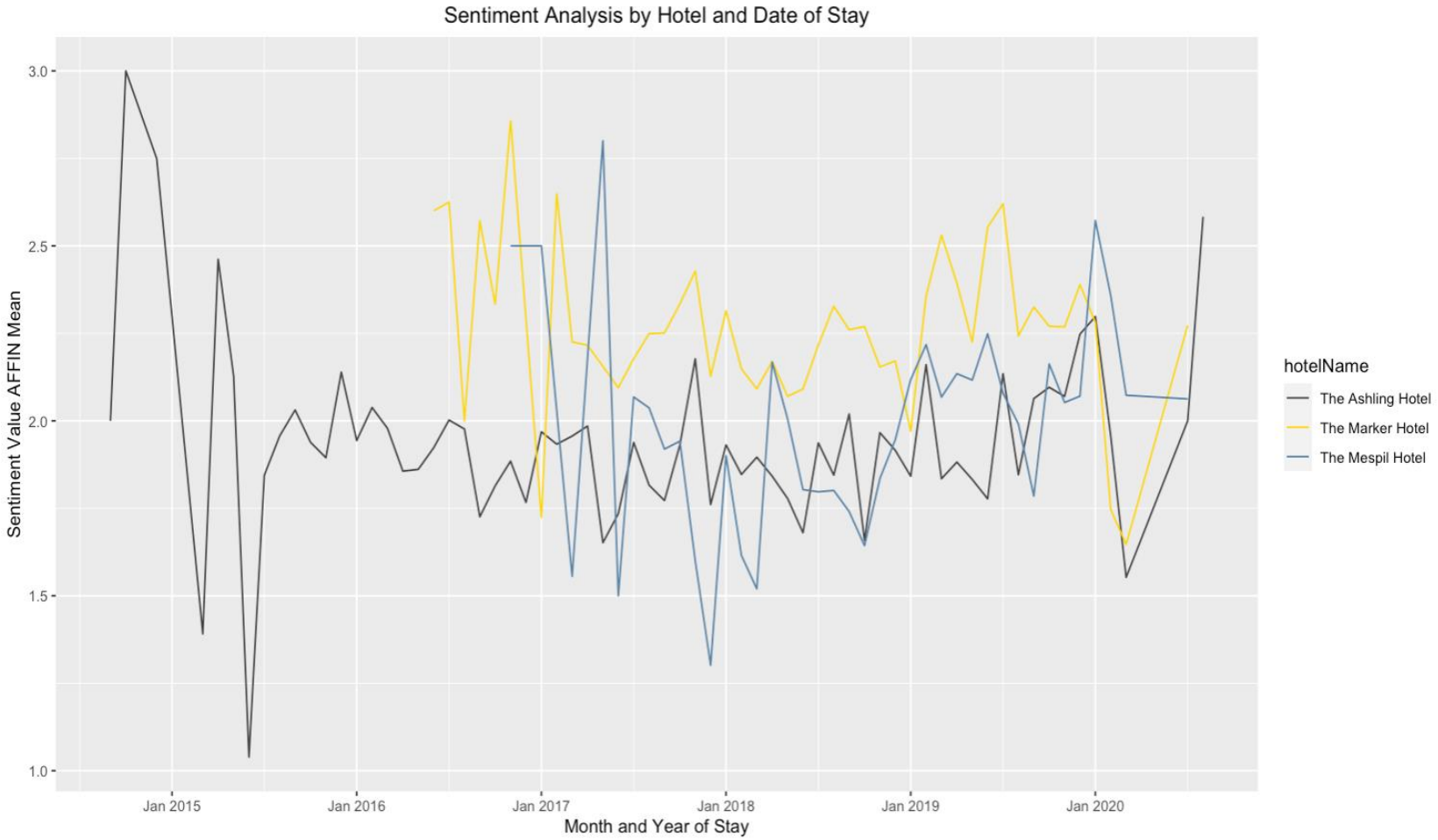


Figure 19: Overall Sentiment Analysis by Hotel – AFINN – Boxplots

By showing the mean over time [Fig.20] , no relevant trend appears other than a big drop in March 2020 for The Ashling and The Marker Hotel.

Figure 20: Sentiment Analysis Mean by Hotel and Date of Stay – AFINN



The second lexicon used was the "Bing" [Fig.22][pag.37, Appendix B].

	hotelName	date	word	sentiment
	<chr>	<yearmon>	<chr>	<chr>
1	The Ashling Hotel	Aug 2020	friendly	positive
2	The Ashling Hotel	Aug 2020	helpful	positive
3	The Ashling Hotel	Aug 2020	spotless	positive
4	The Ashling Hotel	Aug 2020	fantastic	positive
5	The Ashling Hotel	Aug 2020	fantastic	positive
6	The Ashling Hotel	Sep 2019	clean	positive
7	The Ashling Hotel	Sep 2019	spacious	positive
8	The Ashling Hotel	Sep 2019	problem	negative
9	The Ashling Hotel	Sep 2019	promptly	positive
10	The Ashling Hotel	Sep 2019	friendly	positive

… with 22,835 more rows

Figure 21: Ashling Reviews data frame joined with Bing Lexicon

Most of the words have a positive sentiment [Fig.22].

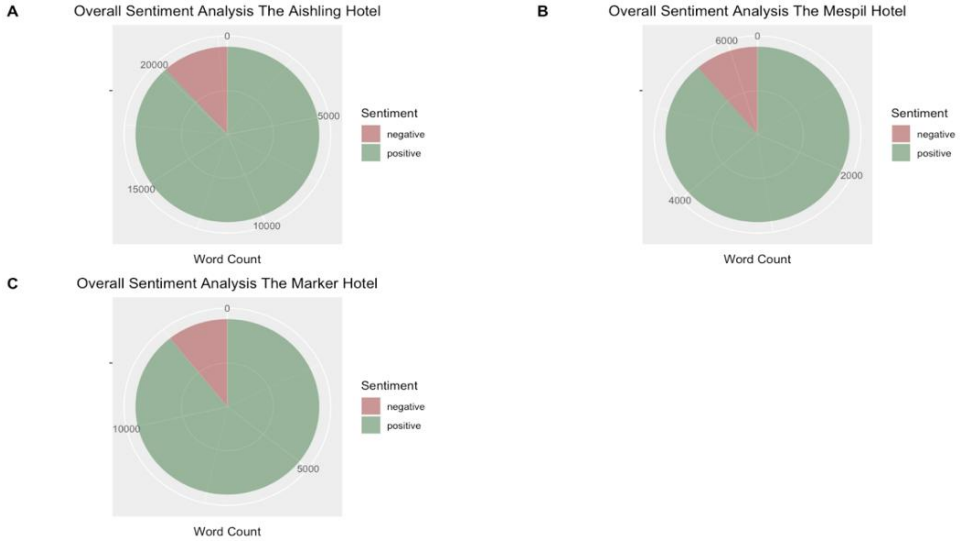


Figure 22: Sentiment Analysis by Hotel and Date of Stay - Bing

By analysing the words contained in the reviews overtime [Fig. 23], it looks like the users tend to reduce the number of words or reviews except for the Mespil Hotel. In all the cases we see a reduction in the latest months probably due to Covid-19.

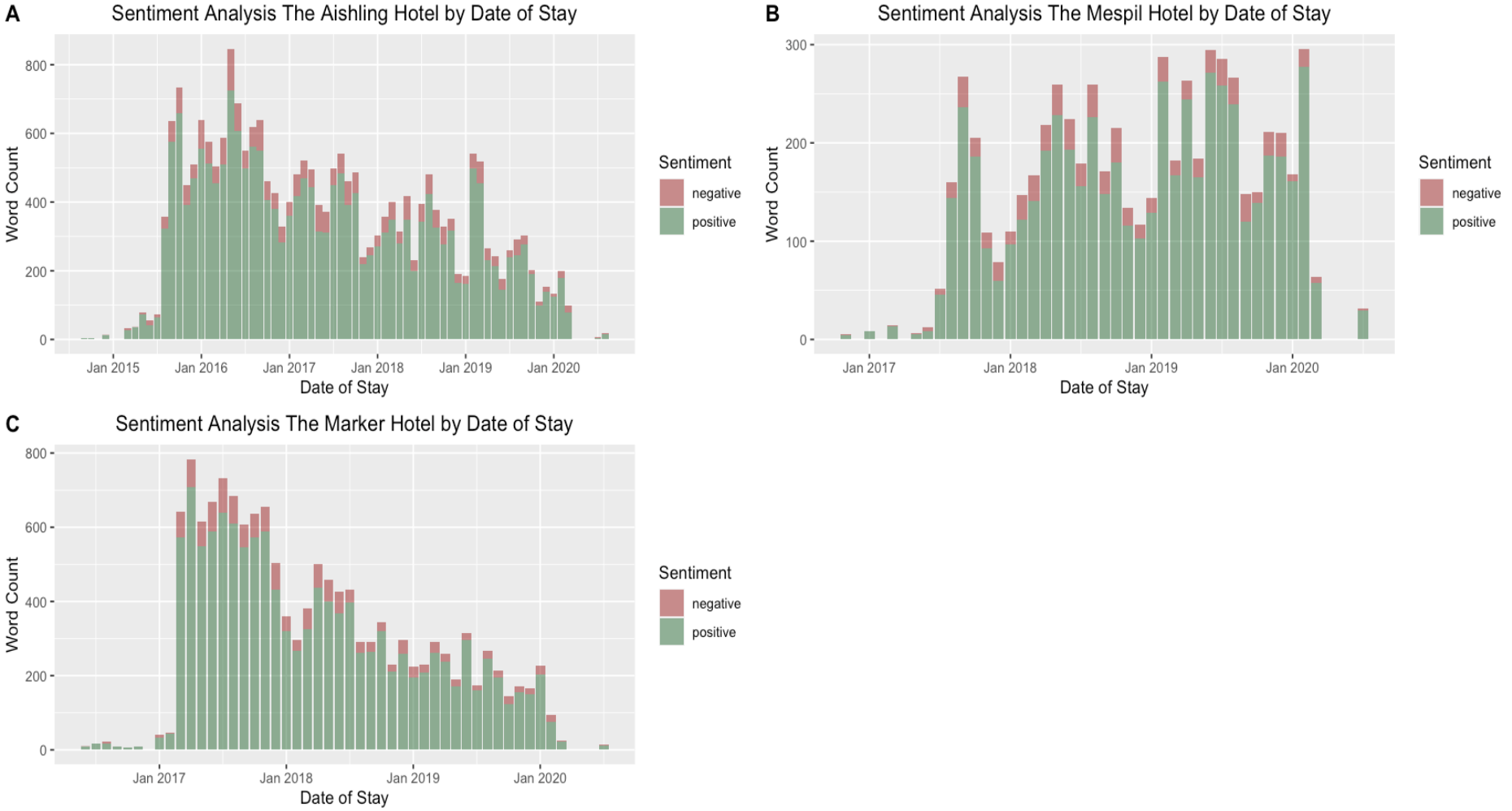


Figure 23: Sentiment Analysis by Hotel and Date of Stay - Bing

When checking the common words associated with positive and negative sentiment [Fig. 24][Fig. 25][Fig.26] it is clear that users like clean, spacious and comfortable hotels with friendly and helpful staff. Some of the most negative sentiment words are “noisy”, “cold” and “expensive”. Regardless of their similarity, the Marker Hotel [Fig. 26] seems to have the word “expensive” stated more times probably due to it being a 5 star hotel.

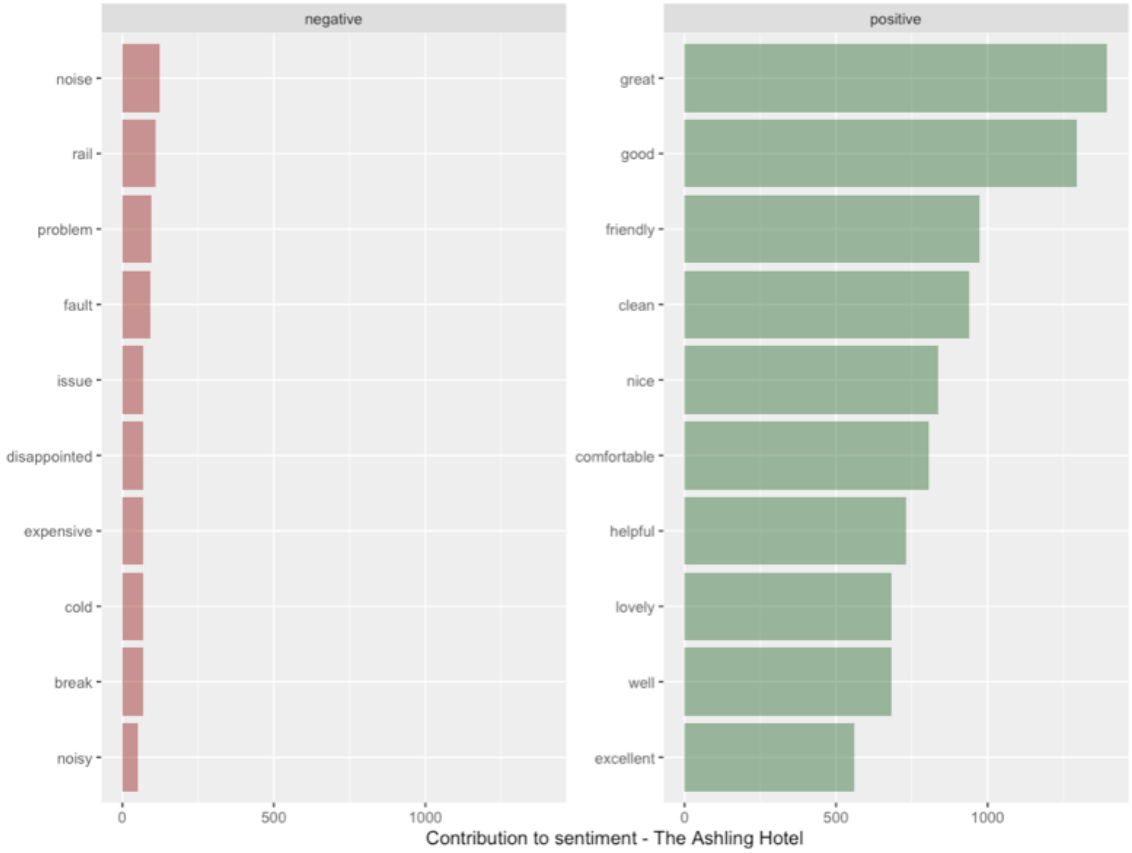


Figure 24: Popular Positive and Negative words The Ashling Hotel

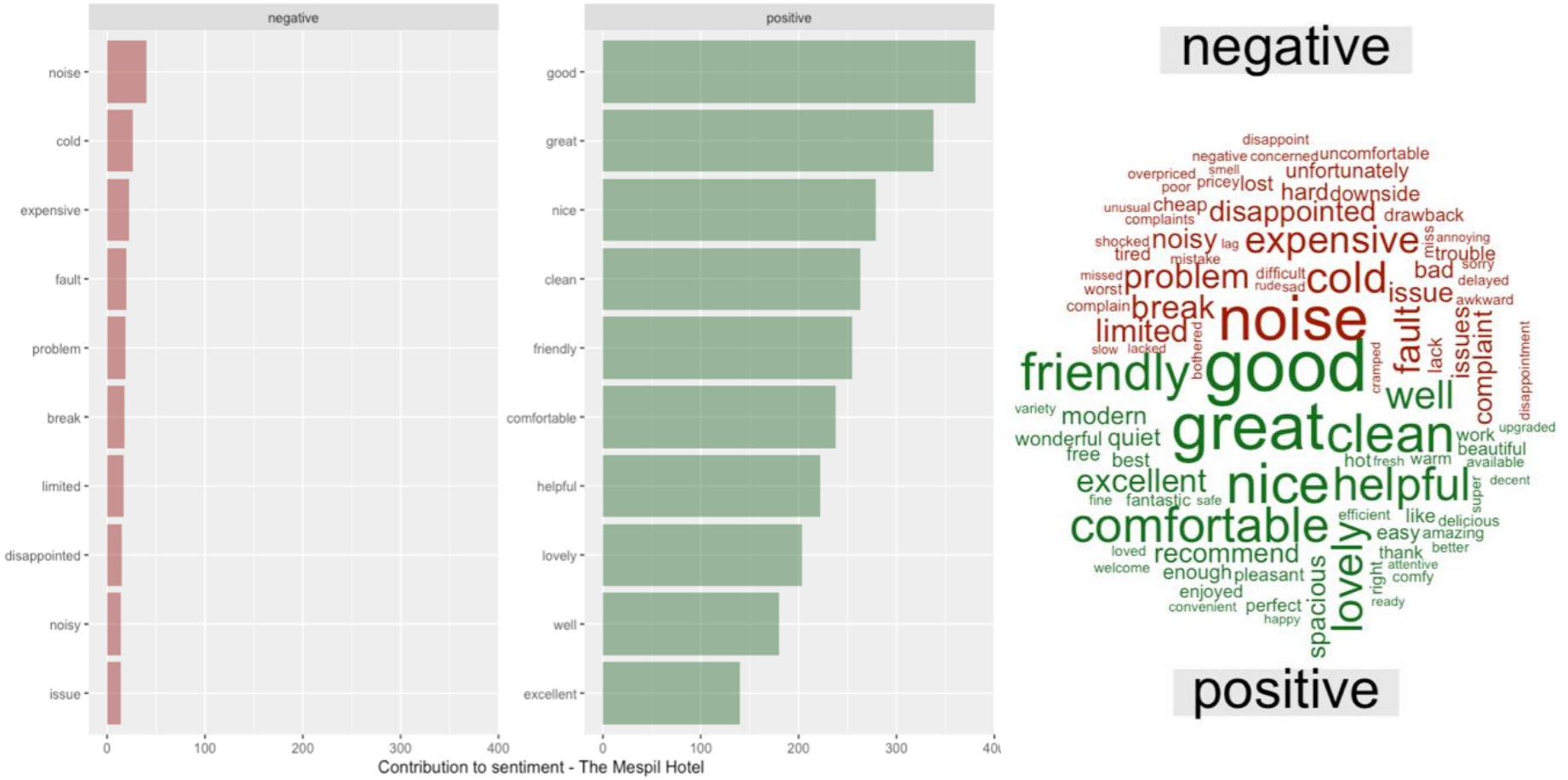


Figure 25: Popular Positive and Negative words The Mespil Hotel

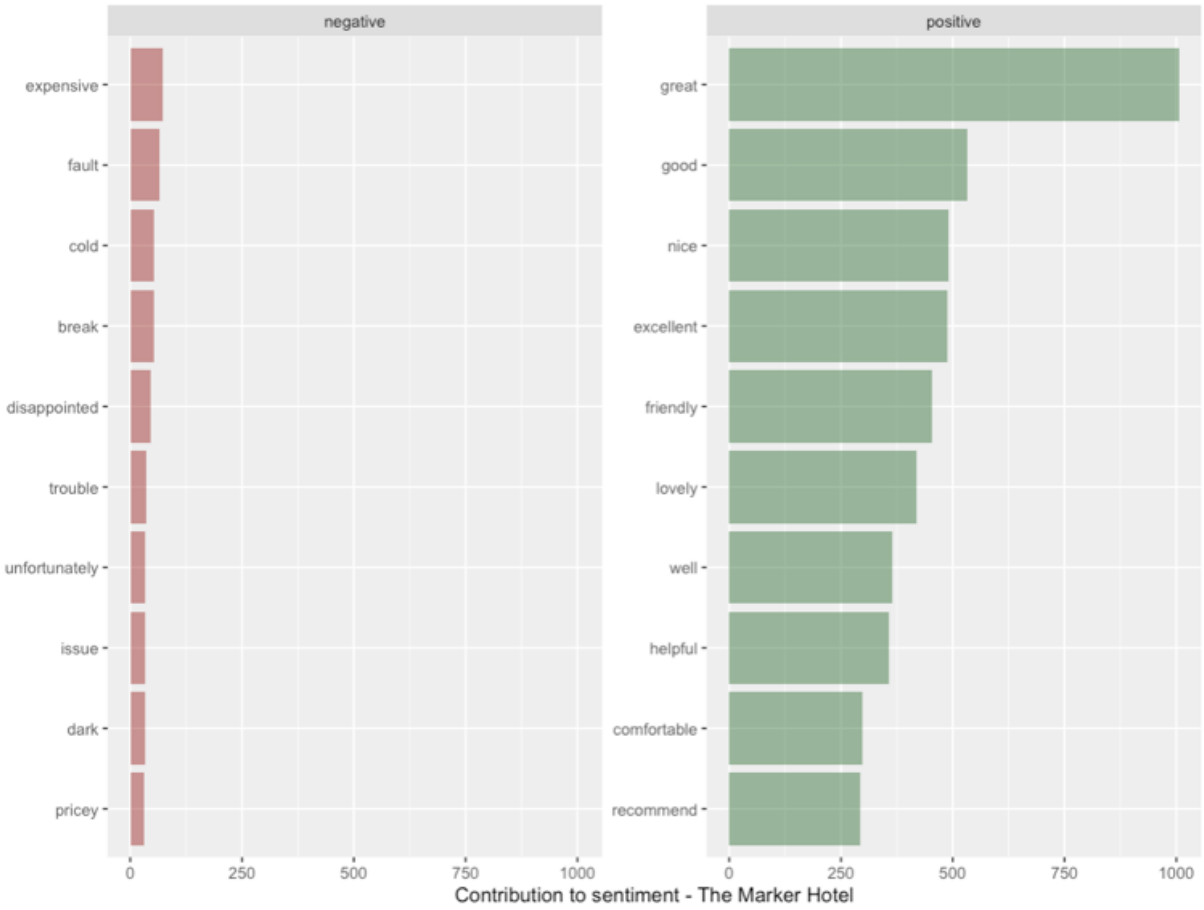


Figure 26 Popular Positive and Negative words The Marker Hotel

The last lexicon used is the “NRC” [Fig. 27] showing again a similar profiles for the three hotels.

	hotelName	date	word	sentiment
	<chr>	<yearmon>	<chr>	<chr>
1	The Ashling Hotel	Aug 2020	friendly	anticipation
2	The Ashling Hotel	Aug 2020	friendly	joy
3	The Ashling Hotel	Aug 2020	friendly	positive
4	The Ashling Hotel	Aug 2020	friendly	trust
5	The Ashling Hotel	Aug 2020	helpful	joy
6	The Ashling Hotel	Aug 2020	helpful	positive
7	The Ashling Hotel	Aug 2020	helpful	trust
8	The Ashling Hotel	Aug 2020	spotless	positive
9	The Ashling Hotel	Aug 2020	spotless	trust
10	The Ashling Hotel	Aug 2020	food	joy

… with 57,173 more rows

Figure 27: Ashling Reviews data frame joined with NRC Lexicon

The most popular positive emotions are “positive”,” trust” and “anticipation”, whereas the negative are “ negative”, “sadness” and “anger”.

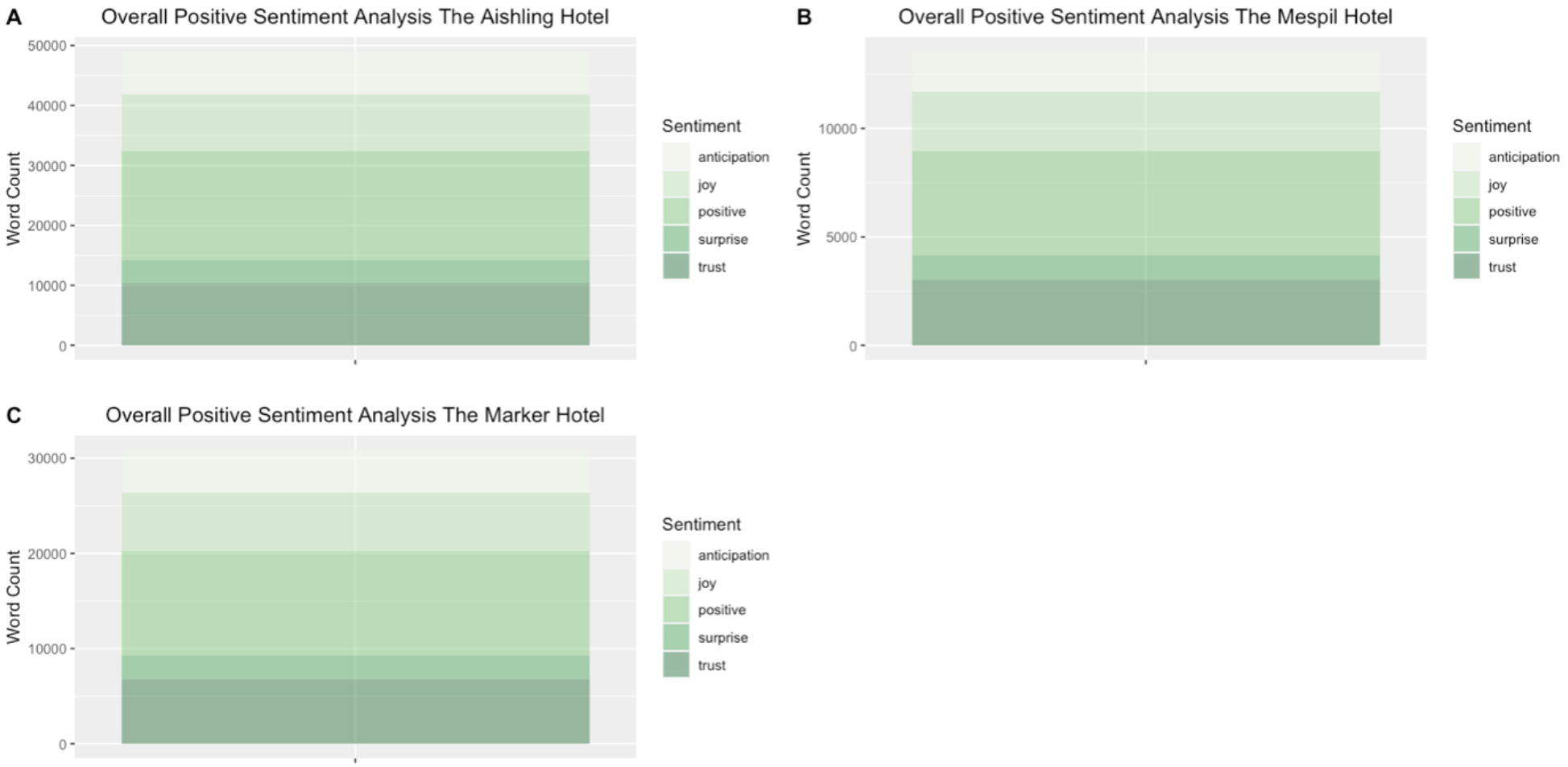


Figure 28: Overall Positive Sentiment Analysis by Hotel – NRC

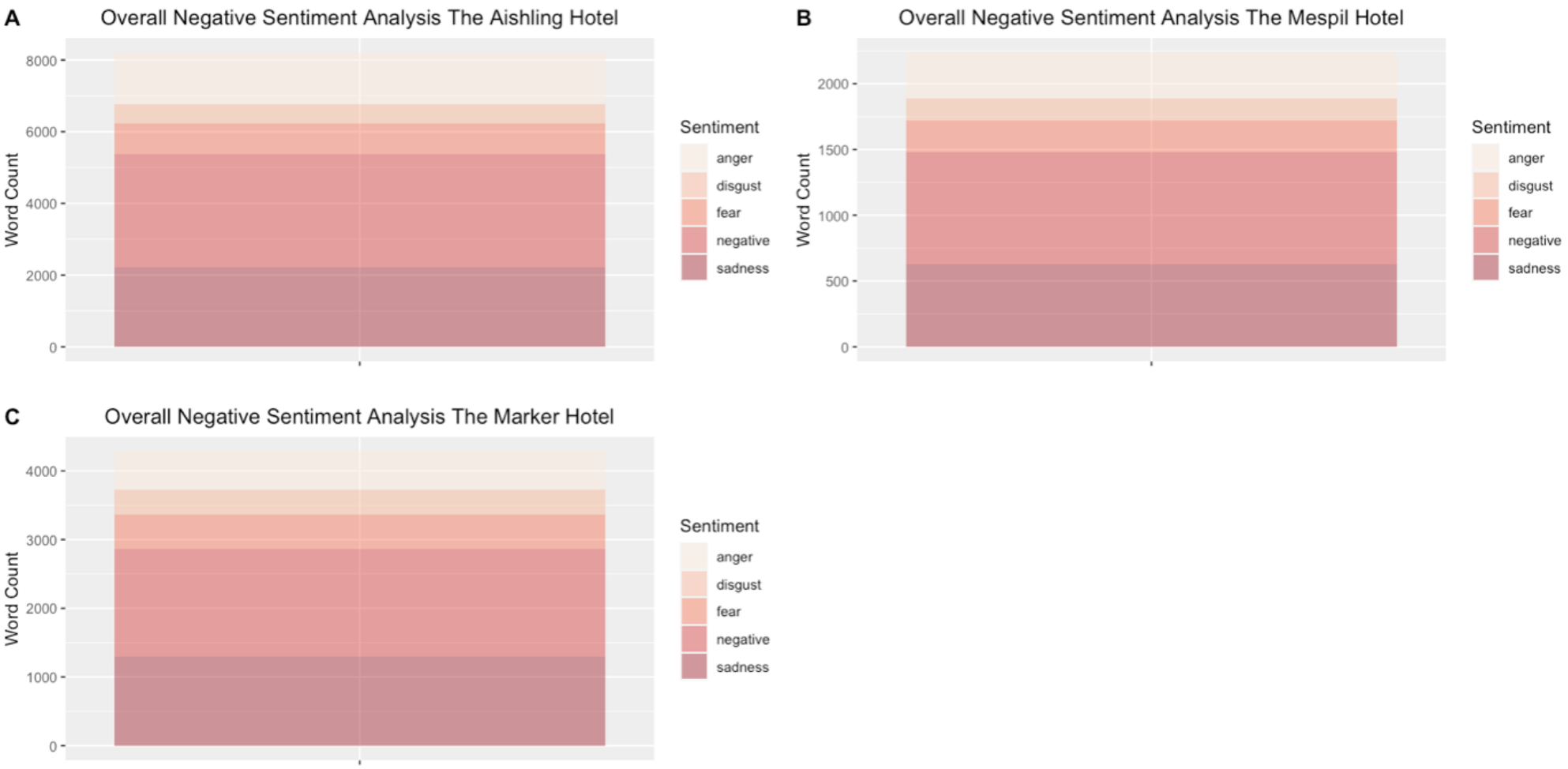


Figure 29: Overall Negative Sentiment Analysis by Hotel - NRC

When looking at the trend over time there seems to be an increase in both positive and negative words [Fig. 30][Fig.31] during the summer months when presumably there are more reservations. The drop due to the COVID-19 is also noticeable together with an increase in the negative, sadness and fear emotions.

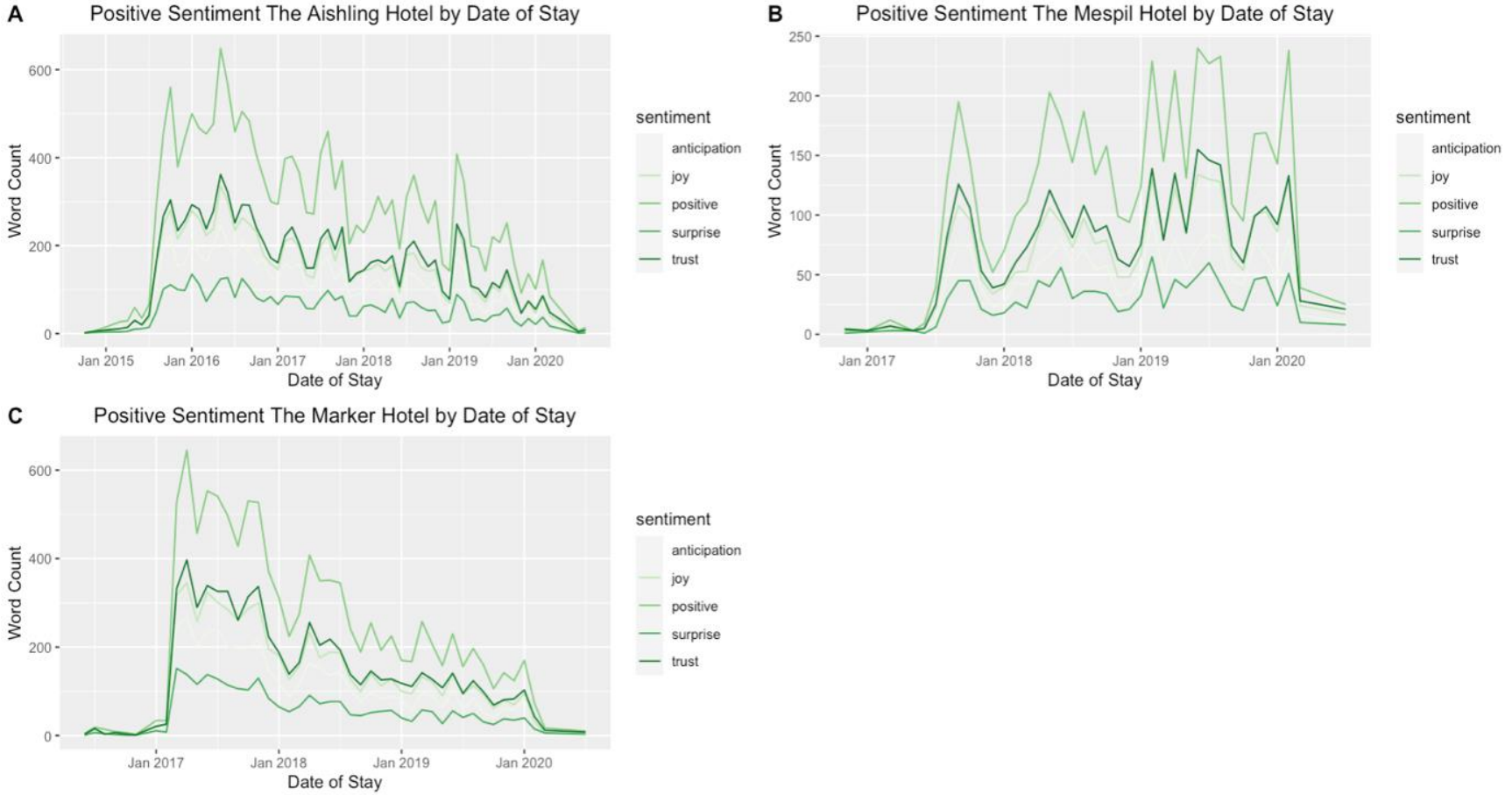


Figure 30: Overall Positive Sentiment Analysis by Hotel and Date of Stay - NRC

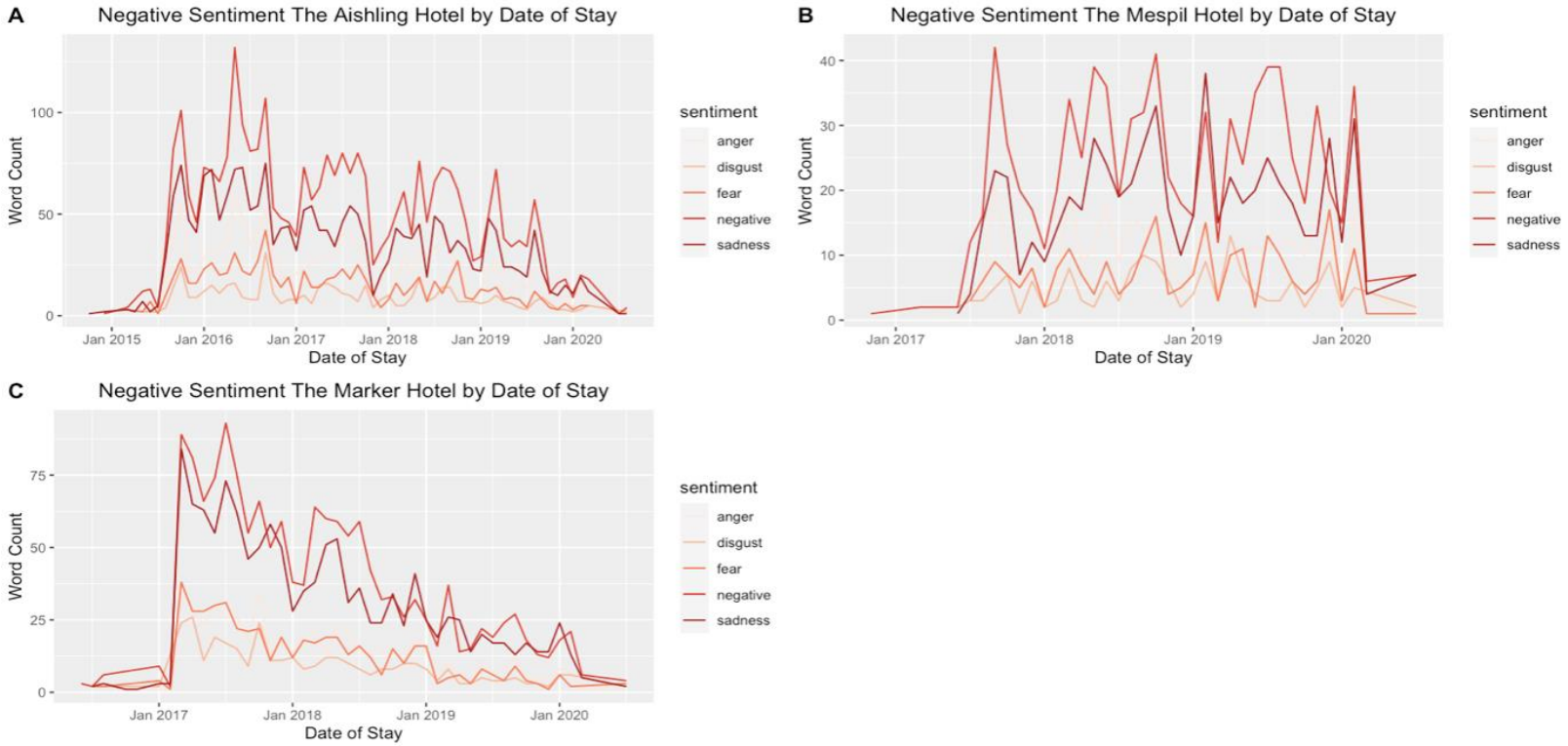


Figure 31: Overall Negative Sentiment Analysis by Hotel and Date of Stay - NRC

The use of lexicon is simple and rewarding but just takes into consideration the words included in the lexicon and takes no account for punctuation or the neighbouring words for example (Naldi, 2019). In some cases analysing the test at sentence or document level, instead of word by word the syntax and grammar can be taken into consideration. Machine learning algorithms (Naïves Bayes, Linear Regression and Support Vector Machine) or Deep Learning could be used for a better result. (Monkey Learn, 2020). In some cases lexicons and machine learning can be combined (rdrr.io, n.d.).

Topic modelling could also be applied to the reviews for each of the hotels separated, to understand which the main topics of interest are.

Conclusion

This analysis began by scraping TripAdvisor.ie for the best-selling hotel within 3, 4, and 5 star hotels in Dublin. This then allowed for the decision to be made that the review and date of stay of each review would be required.

This then was used for the further analysis in the building of the corpus and the creation of the search engine. The creation of the search engine allowed for the performance of 4 queries, one positive query on each hotel and a negative query aimed with a comparison of each hotel that resulted in the Mespil having more negative impactful reviews than the other hotels.

Lastly, a Sentiment Analysis was carried out. The results show that The Marker Hotel received a more negative response in relation to pricing. The Mespil Hotel holds a steadier trend in relation to word count. It can be concluded that overall the three hotels have a similar impact on customers' perception.

Bibliography

- Collaborators, B. L. a., n.d. Opinion Mining, Sentiment Analysis, and Opinion Spam Detection. [Online]
  Available at: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
  [Accessed 18 August 2020].
- Fanara, C., 2018. A Tutorial on Loops in R - Usage and Alternatives. [Online]
  Available at: https://www.datacamp.com/community/tutorials/tutorial-on-loops-in-r
  [Accessed 20th August 2020].
- fnielsen, 2015. AFINN-en-165.txt. [Online]
  Available at: https://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-en-165.txt
  [Accessed 18 August 2020].
- Gabriel, A., 2019. Building a Document Ranking Algorithm in Under 20 lines of Code — in R. [Online]
  Available at: https://medium.com/@ajgabriel_30288/how-to-build-a-search-engine-in-under-20-lines-of-code-in-r-edb79de2febd
  [Accessed 10 August 2020].
- Gabriel, A., 2019. Sentiment Analysis of Political Speeches — Managing Unstructured Text Using R. [Online]
  Available at: https://medium.com/@ajgabriel_30288/sentiment-analysis-of-political-speeches-managing-unstructured-text-using-r-b090a42c0bf5
  [Accessed 10 August 2020].
- Genç, Ö., 2019. The basics of NLP and real time sentiment analysis with open source tools. [Online]
  Available at: https://towardsdatascience.com/real-time-sentiment-analysis-on-social-media-with-open-source-tools-f864ca239afe
  [Accessed 10 August 2020].
- Hadley Wickham [aut, C. and [cph], Rs. (2020) 'Package "rvest"', 0.3.6, pp. 7–8.
  Available at: http://cran.ms.unimelb.edu.au/web/packages/rvest/rvest.pdf
- Jalayer, 2017. Text Mining (part 6) - Cleaning Corpus text in R. [Online]
  Available at: https://www.youtube.com/watch?v=jCrQYOsAcv4
  [Accessed 8 August 2020].
- Khalil, S. and Fakir, M. (2017) 'RCrawler: An R package for parallel web crawling and scraping', 0.1, pp. 8-9(105-106).
  Available at:
  https://www.sciencedirect.com/science/article/pii/S2352711017300110?via%3Dihub
- Lantz, B., 2015. Machine Learning with R. Second ed. Birmingham: Packt. Milborrow, S., 2019. 'Text data in corpus and other packages '. [Online] Available at: https://cran.r-project.org/web/packages/corpus/vignettes/textdata.html [Accessed 18 August 2020].
- Lantz, B., 2015. Machine Learning with R. Second ed. Birmingham: Silge, J., Robinson, D., 2020. Converting to and from Document-Term Matrix and Corpus Objects. [Online] Available at: https://cran.r-project.org/web/packages/tidytext/vignettes/tidying_casting.html
  [Accessed 21 August 2020].
- Lorna,M.A., 2018. Understanding and writing your first Text Mining script [Online] Available at: https://towardsdatascience.com/understanding-and-writing-your-first-text-mining-script-with-r-c74a7efbe30f [Accessed 21 August 2020].

- Mohammad, S. M., 2011. Ten Years of the NRC Word-Emotion Association Lexicon. [Online] Available at: https://medium.com/@nlpscholar/ten-years-of-the-nrc-word-emotion-association-lexicon-eaa47a8dd03e [Accessed 21 August 2020].
- Manning, C. D., Raghavan, P. & Schèutze, H., 2008. An introduction to information retrieval. Cambridge: Cambridge University Press. [Online] Available at: https://library.ncirl.ie/items/19423?query=vector+space+model&resultsUri=items%3Fquery%3Dvector%2Bspace%2Bmodel [Accessed 18 August 2020].
- Monkey Learn, 2020. Sentiment Analysis. [Online] Available at: https://monkeylearn.com/sentiment-analysis/ [Accessed 18 August 2020].
- Naldi, M., 2019. A review of sentiment computation methods with R packages. [Online] Available at: https://arxiv.org/pdf/1901.08319.pdf [Accessed 21 August 2020].
- Nielsen, F. {. 2011. AFINN. [Online] Available at: http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html [Accessed 18 August 2020].
- Nuria Gala, M. L., 2013. NPL lexicons: innovative constructions and usages for machines and humans. [Online] Available at: https://www.researchgate.net/publication/280851799_NLP_lexicons_innovative_constructions_and_usages_for_machines_and_humans [Accessed 21 August 2020].
- rdrr.io, n.d. lexicon_loughran: Loughran-McDonald sentiment lexicon. [Online] Available at: https://rdrr.io/cran/textdata/man/lexicon_loughran.html [Accessed 21 August 2020].
- Robinson, J. S. a. D., 2020. Sentiment analysis with tidy data. [Online] Available at: https://www.tidytextmining.com/sentiment.html [Accessed 10 August 2020].
- SelectorGadget.com (2013) SelectorGadget.com. Available at: https://selectorgadget.com/
- Silge, J. & Robinson, D., 2020. Converting to and from Document-Term Matrix and Corpus objects. [Online] Available at: https://cran.r-project.org/web/packages/tidytext/vignettes/tidying_casting.html [Accessed 21 August 2020].
- Study.com, 2019. Acquiring Data Using readLines Functions in R Programming. [Online] Available at: https://study.com/academy/lesson/acquiring-data-using-readlines-functions-in-r-programming.html#:~:text=The%20readLines%20function%20is%20used,an%20element%20of%20the%20vector. [Accessed 20th August 2020].
- Tatman, R., 2018. Tutorial: Sentiment Analysis in R. [Online] Available at: https://www.kaggle.com/rtatman/tutorial-sentiment-analysis-in-r [Accessed 8 August 2020].
- Tripadvisor, 2020. Dublin Hotels and Places to Stay. [Online] Available at: https://www.tripadvisor.ie/Hotels-g186605-Dublin_County_Dublin-Hotels.html [Accessed 21 August 2020].
- Turney., S. M. a. P., n.d. NRC Word-Emotion Association Lexicon (aka EmoLex). [Online] Available at: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm [Accessed 18 August 2020].

- Weiss, S. M., Zhang, T. & Indurkhya, N., 2010. Fundamentals of predictive text mining. London: Springer. [Online] Available at: https://library.ncirl.ie/items/19427?query=text+mining&resultsUri=items%3Fquery%3Dtext%2Bmining [Accessed 20 August 2020].
- Wickham, H. (2018) str_sub. Available at: https://www.rdocumentation.org/packages/stringr/versions/0.5/topics/str_sub
- Wikipedia, n.d. Vector space model. [Online] Available at: https://en.wikipedia.org/wiki/Vector_space_model#:~:text=Vector%20space%20model%20or%20term,retrieval%2C%20indexing%20and%20relevancy%20rankings [Accessed 21 August 2020].
- Yuan, B., 2017. Sentiment analytics: Lexicons construction and analysis. [Online] Available at: https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=8668&context=masters_theses [Accessed 21 August 2020].
- Available at: https://www.researchgate.net/publication/283954600_Sentiment_Analysis_An_Overview_from_Linguistics [Accessed 23 August 2020].
- Devika M D, S. C. A. G., 2016. *Sentiment Analysis:A Comparative Study On Different Approaches.* [Online] Available at: https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050916X00117/1-s2.0-S187705091630463X/main.pdf [Accessed 23 August 2020].
- ESULI, 2019. *SentiWordNet.* [Online] Available at: https://github.com/aesuli/SentiWordNet [Accessed 23 August 2020].
- Data Monsters, 2017. *Sentiment Analysis Tools Overview, Part 1. Positive and Negative Words Databases.* [Online] Available at: https://medium.com/@datamonsters/sentiment-analysis-tools-overview-part-1-positive-and-negative-words-databases-ae35431a470c [Accessed 23 August 2020].

Appendix A

Tasks Allocation

The Group F contributed equally to this project with regular weekly meeting conducted on Teams to brainstorm the project goals, objectives and tasks to accomplish, as well as the drafting of this document and the video creation. All the members contributed by sharing their R coded scripts on Microsoft Teams to be discussed and presented in the meetings. Below is a detailed subdivision of the members' focus and contribution to the project document;

Section A

Antonio

- Introduction
- Ranking Building Process of the corpus
  - Building Corpus
  -Application of Vector Space Model for Building a Ranking of the Corpus
- Application of queries and interpretation of results.

Jason

- Web Scraping
  - Decision of what to scrape
  -Web Scraping
- Data Cleaning and preparation for Analysis
- Conclusion

Section B

Susana

- Abstract
- Sentiment Analysis
- Sentiment Analysis Lexicons (See Appendix B)

Appendix B

Sentiment Analysis Lexicons

Many lexicons have been created (Yuan, 2017) and many R libraries contains lexicons (Naldi, 2019), but among the most popular are the “Afinn”,”Bing” and “NRC” contained in the “Tidytext” library.

AFINN

```
> get_sentiments("afinn")
# A tibble: 2,477 x 2
   word       value
   <chr>      <dbl>
 1 abandon       -2
 2 abandoned     -2
 3 abandons      -2
 4 abducted      -2
 5 abduction     -2
 6 abductions    -2
 7 abhor         -3
 8 abhorred      -3
 9 abhorrent     -3
10 abhors        -3
# … with 2,467 more rows
```

The AFINN lexicon was created by Finn Arup Nielsen (Nielsen, 2011) . Started including only obscene words and then got extended (Naldi, 2019). Actually contains and includes 2,476 words [Fig.] . Each word has been rated with a value from -5 to 5. Negative values represent negative sentiment and vice versa.

Figure 32: AFINN Lexicon

Bing

The Bing lexicon was created by Minqing Hu and Bing Liu (Collaborators, n.d.) and contain 6,776 words categorized into positive or negative [Fig.].

```
> get_sentiments("bing")
# A tibble: 6,786 x 2
   word        sentiment
   <chr>       <chr>
 1 2-faces     negative
 2 abnormal    negative
 3 abolish     negative
 4 abominable  negative
 5 abominably  negative
 6 abominate   negative
 7 abomination negative
 8 abort       negative
 9 aborted     negative
10 aborts      negative
# … with 6,776 more rows
```

Figure 33: Bing Lexicon

NRC

The NRC lexicon is a bit more complex as contains 10 different emotions, 5 of them on the positive side and 5 on the negative side (Turney., n.d.): (Turney., n.d.)

anger	anticipation
disgust	joy
fear	positive
negative	surprise
sadness	trust

The lexicon was created in 2010 by Saif. M. Mohammad and Peter Turney (Mohammad, 2011). The lexicon contains 13,891 words [Fig.].

```
> get_sentiments("nrc")
# A tibble: 13,901 x 2
   word        sentiment
   <chr>       <chr>
 1 abacus      trust
 2 abandon     fear
 3 abandon     negative
 4 abandon     sadness
 5 abandoned   anger
 6 abandoned   fear
 7 abandoned   negative
 8 abandoned   sadness
 9 abandonment anger
10 abandonment fear
# … with 13,891 more rows
```

Figure 34: NCR Lexicon

Loughran

```
> get_sentiments("loughran")
# A tibble: 4,150 x 2
   word          sentiment
   <chr>         <chr>
 1 abandon       negative
 2 abandoned     negative
 3 abandoning    negative
 4 abandonment   negative
 5 abandonments  negative
 6 abandons      negative
 7 abdicated     negative
 8 abdicates     negative
 9 abdicating    negative
10 abdication    negative
# … with 4,140 more rows
```

Figure 35: Loughran Lexicon

Even if the Loughran lexicon was not used in the analysis, is also include in the "Tidytext" package. This type of analysis is used more for financial sentiment analysis and contain the following sentiments: "negative", "positive", "litigious", "uncertainty", "constraining", or "superfluous" (rdrr.io, n.d.). The lexicon contain 4,140 words [Fig.].