Text Mining and Sentimental Analysis for Online Reviews

Susmitha Chereddy
Swetha Lenkala

Abstract

In the age of internet, reviews have occupied a more prominent place while making any kind of decision. People usually check out for public reviews and research experience from peer groups before making any kind of decisions. These decisions can range from as small as which movie to go for over a weekend to which college/university to choose for their degree. The reviews on the internet can display a wide range of emotions and it usually is a laborious task to go through them before we decide. We are planning to provide a solution to this problem using R while Extracting sentiment out of review using Natural Language Processing.

1 Introduction

Now a days, reviews have become integral part of human life. Almost all the decisions that people take nowadays are subject to reviews from other people. These can be either online or in-person reviews from other people. Getting reviews and perspectives from others is nothing new and has been a norm since the dawn of men. Now that we have internet and people are lot more connected than ever in history, we are able to see a plethora of opinions from people all around the world. This is booth a boon and a curse. It is a boon since we have a lot of information at our disposal that can be used for better decision making. It is curse because of information overload. You need to spend enormous amount to time to read through all the reviews to make out some meaning out of the product/service. Here is where sentiment analysis come in place. Sentiment analysis is defined as a process of extracting sentiment about a review/sentence/entire text. This is usually done by using lexicons that provide us the positive and negative words. These words can be score based on their contributions and frequency which in turn can be used to score the sentence. These sentence score can be further grouped together to get the sentiment for the entire review/text.

Sentiment analysis has huge number of applications in almost every industry including but not limited to e-commere, Hospitality, healthcare, Real-estate, Higher Education and services. Tacking the time-constraints in decision can help us in variety of ways. Hence more focus is need in the field of sentimental analysis to extract a more precise sentiment. Through this paper we plan to taking a stab at the sentimental analysis of user reviews in TripAdvisor on Hilton Hawaiian Village resort in Honolulu.

We are planning to use the following R packages to achieve different functionalities w.r.t the project.

a. dplyr for Data manipulation
b. ggplot2 for Data Visualization
c. tidyr for Data Cleansing
d. reshape2 for data transformation
e. tidytext for analysis and visualization of text
f. purrr for working with functions and vectors
g. scales: to override the ggplot2 defaults.
h. Lubridate: for working with dates and times.
i. widyr: for co-occurrence and co-relations among words
j. snowballC for word stemming

k. wordcloud for analyze and visualize keywords

We plan on utilizing R's data manipulation, transformation and cleansing capabilities combined with text analysis techniques like stemming, lemmatization, stop words, bigrams, and trigrams to achieve at a sentiment for a review. In the process we want to make sure that we preserve the emotion of the review by only removing the stop words from the review. Furthermore, we are planning to visualize the importance and association between different words using graphs, charts, plots, and frequency counts.

2 Data

For this project, we are using the reviews data from Kaggle on Hilton Hawaiian Village Beach Resort located in Honolulu. Data has 13701 rows and 2 columns (review and review date). We will be using R Packages and functionalities to perform text mining and sentimental analysis on this data. Goal of this project is to perform text mining, analyze the words/ emotions contributing to the sentiment of the review and assign a sentiment to each review.

Each row of the data has two columns

a. Date of review
b. Review in words.

We further plan on generating a sequence of numbers which can be used as ID for each review to refer them.

3 Data pre-processing

As a first step of data preprocessing, we only want to select the rows in the data where all columns have data. This can be achieved by complete.cases method in R.

As a next step, we would like create an ID for each row in data. This is done by using tibble which produces a monotonically increasing ID which can be further used to refer each row of the data.

3.1 Removing stop words

Stopwords are defined as connecting words that do not generate any meaning towards the sentence. Most common stop words are "an", "a", "the", "and", "of", "but" etc. These words when removed form a sentence do not change/alter the sentiment of the sentence.

Eg: For a review below,

"We had an excellent stay at the hotel and plan to visit it multiple times in future."

If we remove the stop words from the above review, it would look like below:

"excellent stay hotel plan visit multiple times future"

We see that even after removing the stop words, the sentence give us it's full meaning and sentiment is not changed. Hence, we removed all the stop words and analyzed the frequency of all words in the word corpus to see the frequency of each word.

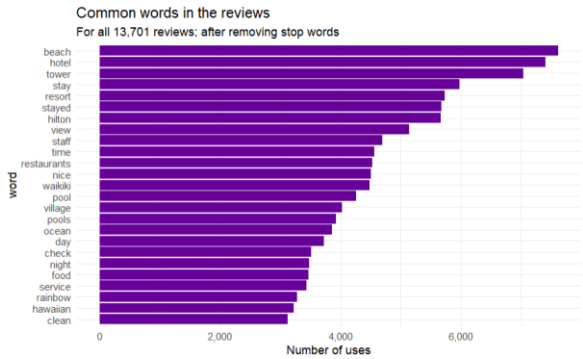


Fig 3.1 Frequency of all the words in the corpus.

3.2 Stemming

Stemming is a process of grouping all the similar words to the stem word while preserving the meaning of the sentence.

Eg: {Played, Playing ,play} => can all be grouped under the stem word play and this would still preserve the meaning of the sentence. After stemming the word corpus, it has modified the frequency count of the different words as shown fig 3.2 below.

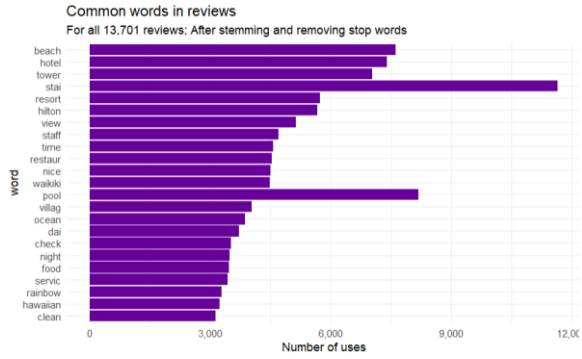


Fig 3.2: Frequency of words altered after stemming

3.3 Bigrams

To understand the sentiment of the sentence better, we need to identify all the words that occur together. This will help us to identify if the sentiment of the sentence changes if they occur together. Hence we extracted all the commonly occurring bigrams in the corpus and made frequency chart out of it.

bigram <chr>	n <int>
rainbow tower	3568
hawaiian village	2909
hilton hawaiian	2821
ocean view	2332
diamond head	2180
waikiki beach	1710
tapa tower	1625
ali'i tower	1583
front desk	1328
resort fee	992

1-10 of 130,453 rows Previous 1 2 3 4 5 6 .. 100 Next

Fig 3.3 Most common bi-grams sin the corpus

We see that most bi-gram is “rainbow tower” which is an attraction point in the resort.

Now, let’s try to generate a word network to see it will help us in making a decision.

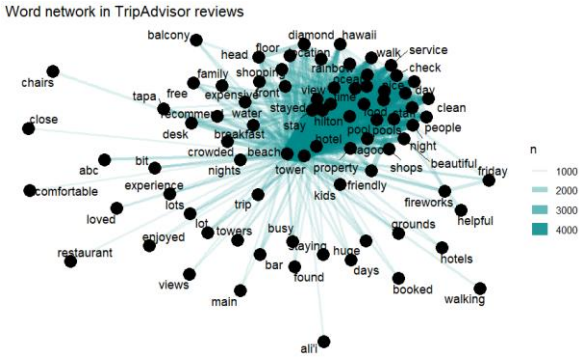


Fig 3.3.1 Word Network of TripAdvisor reviews

Although word network gives us a concentration among certain words, this does not fully tell us about the sentiment of the review/corpus.

3.4 Tri-grams

To further improve on the bi-grams we can analyze set of all three occur together in sentence. A trigram “very very bad” has a greater negative sentiment than “very bad”. Hence let’s extract all the tri-grams that occur in a sentence together and generate a frequency count out of it.

trigram <chr>	n <int>
hilton hawaiian village	2614
diamond head tower	575
partial ocean view	389
ala moana shopping	365
friday night fireworks	358
round table pizza	205
moana shopping centre	171
ala moana mall	147
front desk staff	144
10 minute walk	137

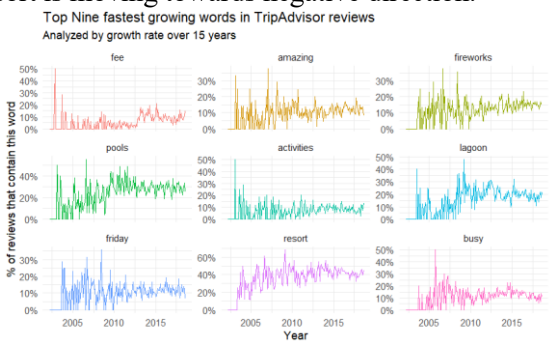
1-10 of 73,678 rows Previous 1 2 3 4 5 6 .. 100 Next

Fig 3.4 Trigrams in the data corpus.

We see that the Hilton Hawaiian village is the most used word since it is the name of the resort.

3.5 Growth/shrinkage Analysis

In this section, we want to analyze the growth/shrinkage of the words over the years. This will help us determine the overall sentiment on the resort. If positive words are growing, which means the sentiment of the resort is growing on the positive side. On the otherhand, if the negative words are increasing, it means the sentiment of the resort is moving towards negative direction.



We see that the fastest growing words here are resort, fireworks and lagoon which does not tell much about sentiment of the entire resort.

Now, let’s look at the words that shrank over the years, to see if they can give us an overall sentiment.

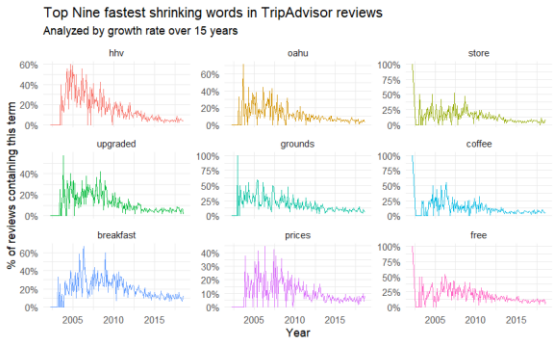


Fig 3.5.2 Top 9 shrinking words

Here, we can see that some positive words like free, breakfast, upgraded have shrunk over the years and this tells us that over all sentiment of the resort is decreasing.

4. Sentiment analysis

Now, let's work on getting a sentiment for each word in the word corpus. For this purpose, we can use sentiment lexicons provided by tidy text package. In the paper, we are discussing two methods:

a. **bing** from Bing Liu and collaborators
b. **AFFIN** from Finn Årup Nielsen

These lexicons provide a positive/negative sentiment to all the words in the English language.

4.1 Bing Lexicon

Bing lexicon divides the words into positive and negative words. It has a repository of 6776 words with a sentiment to each word as shown below.

```
get_sentiments("bing")
#> # A tibble: 6,786 × 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
```

Fig 4.1.0: list of words in bing lexicon with their sentiment.

Now, let's apply bing lexicon to all the words in the word corpus and see the frequency of the negative and positive words.

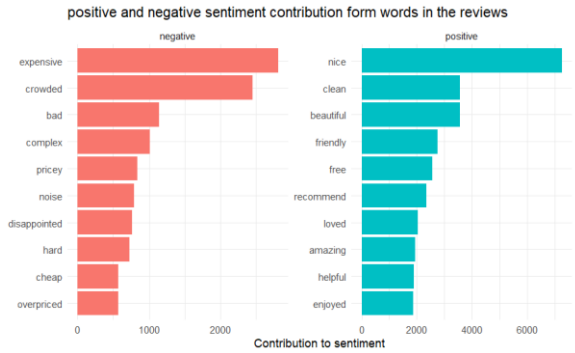


Fig 4.1.1 Frequency of negative and positive after bing lexicon.

We can see that the negative word "expensive" have occurred more than 2000 times and top positive word is nice which occurred more than 7000 times.

4.2 AFFIN Lexicon

While bing lexicon is binary and assigns -1/+1 to all the words, AFFIN Lexicon on the other hand assigns sentiment in the range of -5 to +5

```
get_sentiments("afinn")
```

```
#> # A tibble: 2,477 × 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
```

Fig 4.2.0 list of words in AFFIN Lexicon and their sentiment.

From fig 4.2.0, We can see that AFFIN gives sentiment in the range of -5 to +5.Now, let's apply AFFIN to the word corpus to get the sentiment and the frequency.

word2 <chr>	value <dbl>
worth	2
recommend	2
like	2
want	1
great	3
good	3
bad	-3
happy	3
impressed	3
disappointed	-2

1-10 of 252 rows Previous 1 2 3 4 5 6 ... 26 N

Fig 4.2.1 Word frequency with sentiment after AFFIN.

Here, we can see that AFFIN assigns sentiment to the words based on the strength of the words and how they contribute to a sentence.

Since AFFIN has a relative sentiment score, it is very useful when we are trying to get sentiment from a huge data/text corpus.

4.3 Sentiment of a review

Now, let's apply bing lexicon and calculate the sentiment of each review in our data.

After apply bing lexicon, we looked at the top rated reviews ordering by the sentiment score.

ID <int>	sentiment <dbl>	words <int>
2363	4.000000	7
2578	3.800000	5
3629	3.800000	5
13692	3.800000	5
1039	3.714286	7
1332	3.714286	7
12954	3.714286	7
4909	3.666667	6
5768	3.666667	6
11573	3.666667	6

From the data we see that review 2363 has the highest positive sentiment score and let's see if our analysis is right.

```
data1[ which(data1$ID==2363), ]$review_body[1]
```

```
[1] "Wow wow wow what a place we had a fantastic time and amazing views from our room. The Hilton is just fantastic This Hot
el is amazing for everyone from kids to adults it has everything you will ever need."
```

Review 2363: [1] "***Wow wow wow*** *what a place we had a* ***fantastic*** *time and* ***amazing*** *views from our room. The Hilton is just* ***fantastic*** *This Hotel is* ***amazing*** *for everyone from kids to adults it has everything you will ever need.*"
Definitely looks like a positive review and there are seven words(wow,wow,wow,fantastic,amazing,fantastic, amazing) that have contribute to this sentiment.

Now, let us look at top rated negative review to see if our analysis makes sense:

ID <int>	sentiment <dbl>	words <int>
3748	-3.60000000	5
317	-2.80000000	5
7135	-2.60000000	5
10498	-2.60000000	5
2766	-2.28571429	7
10656	-2.28571429	7
24	-2.20000000	5
9161	-2.20000000	5
12080	-2.20000000	5
692	-2.16666667	6

1-10 of 8,515 rows Previous 1 2 3 4 5 6 ... 100 Next

Fig 4.3.1 Top negative review.

From fig 4.3.1, we can see that top rated negative is review ID 3748.

Review 3748: *"[1] "Stayed here for 5 nights 5/12/16 - 5/17/16. The first night we noticed one of the floor tiles broke and kids were playing with their fingers. The second night we start seeing small cock roaches crawling on our baby's food. The front desk offered new room however they request is to move within 1 hour, otherwise they can't let us switch... It was late in the night 11pm, we all tired and babies were already sleep. We decline the offer. When checkout, the front counter lady told me cock roaches are common in this hotel rooms. And asked me California don't have any cock roaches? This is way beyond my expectation for Hilton.*"

After going through this review, we can say that this a negative review.

5. Future Extension

In this paper, we have discussed about sentiment extraction from text. However, in a growingly digital world, we can have data in multimedia (audio/image/video) format.

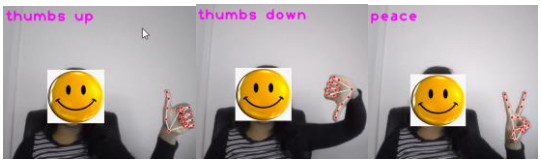
If the data is an audio, we can use the speech to text converter to convert audio to text and then use the similar text mining techniques to generate sentiment for the review.

If the data is an video/image format, we can make use of opencv and mediapie to recognize the gesture and extract sentiment out of it.

First let's generate some hand gestures to see if we can recognize them.

Now let's try and see if we can recognize the hand in the gesture using the mpHands Functionality.

Now, let's see if we can make out the hand gestures using a trained tensorflow model and predict them.



We see that the model predicts hand gestures as needed.

6. Conclusion

We have made a decent attempt at solving the problem of extracting sentiment from the reviews. Also, we have expanded this to video format of the data to extract the sentiment. This can further expanded with speech and video to dynamically generate sentiment from the multimedia review.

References

P. Andersen, What Is Web 2.0? Ideas, Technologies and Implications for Education, tech. report, JISC Technology & Standards Watch, 2007; www.ictliteracy .info/rf.pdf/Web2.0_research.pdf.

E. Cambria, "Affective Computing and Sentiment Analysis," IEEE Intelligent Systems, vol. 31, no. 2, 2016, pp. 102–107.

B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge Univ. Press, 2015.

World Travel and Tourism Council, Travel & Tourism Economic Impart 2016 World, 2016; www.wttc.org//media/files/reports/economic%20 impact%20research/regions%202016/world2016.pdf.

P. O'Connor, "User-Generated Content and Travel: A Case Study on TripAdvisor.com," Information and Comm. Technologies in Tourism, P. O'Connor, W. Höpken, and U. Gretzel, eds., Springer, 2008, pp. 47–48.

J. Serrano-Guerrero et al., "Sentiment Analysis: A Review and Comparative Analysis of Web Services," Information Science, vol. 311, Aug. 2015, pp. 18–38.

E. Cambria and A. Hussain, Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis, Springer, 2015.

F.H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter Opinion Mining Framework Using Hybrid Classification Scheme," Decision Support Systems, vol. 57, Jan. 2014, pp. 245–257.

E. Guzman and W. Maalej, "How Do Users Like This Feature? AFine Grained Sentiment Analysis of App Reviews," Proc. IEEE 22nd Int'l Conf. Requirements Eng., 2014, pp. 153–162.

D. Gräbner et al., "Classification of Customer Reviews based on Sentiment Analysis," Information and Comm. Technologies in Tourism, Springer, 2012, pp. 460–470.

S. Palakvangsa-Na-Ayudhya et al., "Nebular: A Sentiment Classification System for the Tourism Business," Proc. 8th Int'l Joint Conf. Computer Science and Software Eng., 2011, pp. 293–298.

K. Ilieska, "Customer satisfaction index-as a base for strategic marketing management", *TEM Journal*, vol. 2, no. 4, pp. 327, 2013.

Abdullah, T. (2017). Penilaian Wisatawan akan Atribut Pariwisata di Kota Batu. THE Journal : Tourism and Hospitality Essentials Journal, 7(2), 91.

Bandur, A. (2016). Penelitian Kualitatif : Metodologi, Desain, dan Teknik Analisis Data Dengan Nvivo 11 Plus (1st ed.; Jatmiko, ed.). Jakarta: Mitra Wacana Media.

Web Scraping TripAdvisor, Text Mining and Sentiment Analysis for Hotel Reviews | by Susan Li | Towards Data Science

Real-time Hand Gesture Recognition using TensorFlow & OpenCV - TechVidvan

Fast-Tracking Hand Gesture Recognition AI Applications with Pretrained Models from NGC | NVIDIA Technical Blog

M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” Proc. 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, 2004, pp. 168–177.

K. Shouten and F. Frasincar, “Survey on Aspect-Level Sentiment Analysis,” IEEE Trans. Knowledge and Data Eng., vol. 28, March 2016, pp. 813–830.

M. Atzmueller, “Subgroup Discovery,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 1, 2015, pp. 35-49.

S. Poria et al., “A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks,” Proc. 26th Int’l Conf. Computational Linguistics (COLING16), 2016, pp. 1601–1612.

F. Herrera et al., Multiple Instance Learning: Foundations and Algorithms, Springer, 2016.

Supplementary Material

https://www.kaggle.com/code/kerneler/starter-hilton-hawaiian-village-048bdbd4-9/data

https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

https://www.kaggle.com/code/wiktorbrk/trip-advisor-reviews-sentiment-analysis

https://www.kaggle.com/code/frankschindler1/sentiment-analysis-tripadvisor-reviews

https://www.kaggle.com/code/mmaguero/tripadvisor-sentiment-analysis-for-hotel-reviews/output

https://www.kaggle.com/code/wiktorbrk/trip-advisor-reviews-sentiment-analysis/comments

https://www.kaggle.com/code/mmaguero/tripadvisor-sentiment-analysis-for-hotel-reviews/comments

https://www.kaggle.com/questions-and-answers/142233

https://www.kaggle.com/code/jonathanoheix/sentiment-analysis-with-hotel-reviews