Brandon Liunoras

IST 707

Final Project

Table of Contents

[Introduction 3](#_Toc138080963)

[Data and Problem 3](#_Toc138080964)

[Analysis 4](#_Toc138080965)

[Models 5](#_Toc138080966)

[Conclusion/Actions to Take 13](#_Toc138080967)

# Introduction

Data mining and business go hand in hand in today’s world. With the rise of large corporations such as Amazon and Google, it is important to find a way to understand the feelings of end users without having to go through data by hand. The method of reading and analyzing customer sentiment individually would be painstakingly long and would not provide actionable insight within a timely manner. However, with sentiment analysis and algorithms, data can be more easily and quickly analyzed to create solutions.

Sentiment analysis, also referred to as opinion mining, is a type of analysis that can both quantify speech and text into numerical scores for interpretation. These scores are made by using algorithms and machines that interpret inflection and word usage within speech. For text data, the data is put through an algorithm that then produces a combined score to tell if the overall sentiment was positive or negative. Other algorithms can also be used to find the most common used words within the reviews as well as finding specific emotions such as joy and fear based on word combinations.

# Data and Problem

Within this project the data that is used is training data reviews that consists of user comments of a resort that I previously worked at. The reviews were made regarding client stays between July through September of 2022. All the data has no mention of personal information such as names, contact information, or proprietary information.

The issue that is looking to be solved is to find the causes of why overall resort revenues decreased between July 2022 through September 2022. Using the resort review data stated before, sentiment analysis can be used on the reviews to understand customer sentiment and see if there were any issues that guests experienced. After understanding the consensus of the guests, actionable insight will be gained, and new solutions can then be created in order to fix the problems that were analyzed from the reviews.

# Initial Analysis/Methodology

When initially looking at the data, the data is already clean with 857 rows and 4 different columns. The four columns go as follows: “review\_id”, “rating”, “comment”, and “is\_negative”. The first column is a unique column that represents the ID of a review. The second column is a user input column with numbers ranging from 1 through 5. These numbers provide a rating of their review with 1 being the worst score and 5 being the best score. The third column is a text string column that contains the text of each review created by a guest. The last column is a factor column that shows whether the review is positive or negative. If the column displays a “1”, the review is negative in sentiment. If the column displays a “0”, the column is neutral or positive in sentiment. The factor of whether a comment was negative or not was based on the rating column. If the rating was “1” or “2”, the review would be classified as negative or “1”. If the rating was “3”, “4”, or “5”, the review was classified as positive/neutral or “0”.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1: The first five rows in the original dataset

The main method that is used in this project to create sentiment analysis is to use several algorithms that provide polarity scores to the comment. The three different vectors that are used are the afinn vector, the bing vector, and the syuzhet vector. They differ in scale, but all provide the same overall function which is to provide a polarity score based on the review sentiment presented solely in the text comment. After the initial analysis, word counts are taken in order to find different trends among all the reviews provided by the guests.

# Models

To start, the comments are separated by their classification of whether they are positive or negative. The comments are then passed through to the different vectors algorithms to provide polarity scores. Even though with different scales, each vector shows the same pattern depending on if a comment is positive or negative. Most of the positive reviews have polarity scores that are greater than 0 and the negative reviews have polarity scores that are negative and towards 0. In the next six figures, the polarity score density is shown between each vector and between the positive and negative reviews.

A picture containing text, diagram, plot, line

Description automatically generatedA picture containing text, plot, diagram, line

Description automatically generatedA picture containing text, plot, diagram, line

Description automatically generatedA picture containing text, plot, diagram, line

Description automatically generatedA graph with a green line

Description automatically generated with low confidenceA picture containing text, plot, diagram, line

Description automatically generated

After looking at the polarity scores, term document matrices are created for all reviews combined, and positive and negative reviews separately. A word count algorithm is then implemented to analyze what the most popular words are within each dataset. Within all reviews, the most popular words are “room”, “staff”, and “great”. This, however, accounts for all the reviews. Among the positive reviews, the most popular words are “great”, “staff”, “hotel”. Among the negative reviews, the most popular words are “room”, “spa” and “clean”. The next few figures display the top five most frequent words in the form of both table count and bar graphs. A picture containing text, screenshot, diagram, display

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generated

Along with the word charts and tables, wordclouds are created to show not only the main trends, but smaller topics that were discussed within the reviews themselves. Within the positive wordcloud, other words that are included are “beautiful”, “beach” and “pool”. From this information, it can be concluded that the pool and beach aspects of the resort are well enjoyed by guests. The negative wordcloud had the words “unclean” and “filthy”. This allows for inferences to be made that certain guests believed the resort did not attain a certain level of cleanliness. The next two figures show the positive and negative review wordclouds respectively.

A close up of words

Description automatically generated with low confidence

A picture containing text, font, design

Description automatically generated

Lastly, a sentiment chart was created to get the different range of emotions that were displayed within the comments themselves rather than a score. The NRC Emotion Lexicon was used for this analysis. The NRC Emotion Lexicon is a library of English words that are associated to an emotion. To display an emotion based on a comment, the comment is first put into the lexicon. The lexicon then analyzes the words within the comment and then matches each word with its corresponding emotion. Once the comment is analyzed, an emotion is displayed based on the word-emotion count, which shows the overall emotion of the comment. In the next figure, joy and trust appear to rank as the two most experienced emotions within the reviews. However, there are another range of emotions that include anger, sadness, and disgust. The next two figures display the sentiment table count and the sentiment review bar graph.

A screenshot of a black screen

Description automatically generated with low confidence

A picture containing text, screenshot, diagram, plot

Description automatically generated

# Conclusions/Actions to Take

Overall, with the sentiment analysis techniques used on the dataset, new information has been gained in a relatively quick manner and has provided meaningful insight. The data set shows that more than three quarters of the resort goers are happy and had a positive experience with the resort. Most of these guests talked about the beauty of the resort and that the staff had a positive experience based on the wordcloud information. The emotional chart also revealed that most reviews from guests show some form of joy and trust. This goes to show that the resort should continue to put in the same effort to maintain positive reviews and maintained revenues.

However, it is seen that there is still 23% of the resort goers had negative experiences. The main things discussed in these reviews are the spa and rooms. This means that action should be taken to investigate the hotel rooms and the spa. Based on the wordcloud, it is important to note the cleanliness of the spa and rooms as the word “clean” was within the negative wordcloud. Taking these actions are only the first steps into increasing revenues from a client-based focus.

Sentiment analysis is an important tool within the data field. With its capabilities to take in large amounts of text data and audio data, it allows for anyone using it to find out the emotion of the end users of any service or product that is being used. From a real-world perspective, any company, big or small, can use sentiment analysis in improvements and maintenance in any field.