Brandon Liunoras

IST 718

Final Project

Table of Contents

[Introduction 3](#_Toc138080963)

[Data and Problem 3](#_Toc138080964)

[Analysis 4](#_Toc138080965)

[Models 5](#_Toc138080966)

[Conclusion/Actions to Take 10](#_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”.

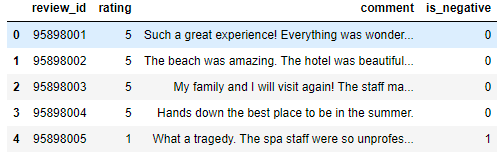


Figure : The first five rows in the original dataset

The main methods that were used for sentiment analysis include the Natural Language Toolkit (NLTK), the Sentiment Intensity Analyzer, and a Random Tree Classifier. The Sentiment Intensity Analyzer is from the NLTK package and will be used to provide the polarity scores of the reviews. The Random Tree Classifier will be used in order to verify the actual sentiment of the review made by the guest. It is already denoted within the dataset if a comment is negative. However, it is important to note that there sometimes may be disconnect between the rating and the comment, potentially making the last column disingenuous. This makes it important to find the true sentiment score of the comment itself and not focusing anything on the rating.

# Models

The first step to understanding the true sentiment of the as stated before is to use the Sentiment Intensity Analyzer. In order to use it, the comment must be cleaned and stripped of any numbers, punctuation, and “stop words”. Stop words are words that are frequently used prepositions, subject articles, and other words that do not provide any additional information to the set. Such words include “is”, “a”, “the”, and many more. The cleaning then provides a new text column with the words that were not stripped out of the initial comment. After the initial text comment is cleaned, the new text column is then inserted into the Sentiment Intensity Analyzer and provides four different scores. Each score shows how strong the positivity, neutrality, and negativity of each comment is. Those three scores are then used to create a compound score to create the overall score and sentiment.

A screenshot of a computer

Description automatically generated with low confidence

Figure : New dataset with added cleaned comment column and polarity scores.

As for the scores themselves, the positive reviews had polarity scores mainly between 0.7 and 1.0 whereas the negative reviews had polarity scores between -1.0 to 0.0. The greater the distance the score is from 0, the greater the intensity of the polarity of the review. The distribution between positive/negative reviews is that the positive reviews make up 77% of all reviews and the negative reviews make up 23% of all the reviews. This means that a large population of the reviews have strong positive sentiment towards the resort. However, there is still a sizeable amount of reviews that are negative that can help uncover as to why there are lessened resort revenues for the summer.

A picture containing diagram, plot, line

Description automatically generated

Figure : Polarity Score Density Graph of the Positive Reviews

A picture containing diagram, plot, line, text

Description automatically generated

Figure : Polarity Score Density Graph of the Negative Reviews

The second step to the analysis is to see if the polarity scores match up to the classification that they were initially given. Doing this will show if the guests’ ratings provide the same sentiment as the comments that they provided. The accuracy tests that were used to prove if the positive/negative classifications are correct, were the Receiver Operator Characteristic and Precision-Recall Curve Tests. For the tests themselves, their accuracy was 0.98 out 1.00 and 0.95 out of 1.00 respectively. This means that the ratings and comments will almost always go hand in hand and that the guest reviews are true with their rating and review together.

A picture containing text, line, plot, diagram

Description automatically generated

Figure : ROC Accuracy Test for the Classification of Positive/Negative Reviews -- 98% Correct

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Figure : PR Accuracy Test for the Classification of Positive/Negative Reviews -- 95% Correct

Lastly, wordclouds are used to see what are the most frequently said words that within the reviews. The two wordclouds that were created are wordclouds where the negative and positive reviews were separated. This separation of reviews shows what each group is saying and find the trends between the two. Within the positive review wordcloud, the main words to note include the words “great”, “staff”, and “hotel”. As for the negative review wordcloud, the words “room”, “spa”, and “clean” are the most prominent within the wordcloud.

From the positive review wordcloud it can be concluded that the staff are positive factors within guest stay and that the hotel itself seems to be positive in some way shape or form. Based on other words within that wordcloud, the words “beautiful”, “amazing”, and “beach” are listed. This shows that guests like the ambiance of the resort. With the negative review wordcloud, it can be concluded that the hotel rooms and the spa are suspected in providing a negative experience to the guests and patrons. Along with the word “clean”, the word “mess” is also listed within the wordcloud. This can provide some insight into the resort that there are certain areas that are not up to cleanliness standards to the guests.

A close-up of words

Description automatically generated with low confidence

Figure : Positive Review Wordcloud showing the most frequent words

A close-up of words

Description automatically generated with low confidence

Figure : Negative Review Wordcloud showing the most frequent words

# 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. 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.