Name: **Kurt Jurgen Von D. Lumbao** Year/Course: **1/MIS**
Date: **August 30, 2022** Score:

OVERVIEW

In this presentation, there is an integration of large data sets in which can be useful in the operational aspect of every organization where decision makers can oversee the potential outcome and impact of its features and processes. Data can be processed into useful information where consolidated information will create solutions and concrete analysis on how to run a certain feature into more efficient and effective.

Nowadays, we have a lot of available resources which are open source while we are on the organization considering the cost of the mechanisms as an emblem to fast track our solutions to our clientele. The interpretation of data that will be done manually will take time and also limited. Like, a hundreds of data has been recorded, cannot be interpreted in one look. Thus, data science has a great contribution to the technology where it enables the interpretation made easy.

Many of these programming platform has evolve since it has began. Like generations to generations it was really upgraded. In most time, some empirical reasons why software and hardware developers are shifting from the mean time process to the new process and it is really the cycle. Python has become one of the most powerful platform and mechanism in interpreting large amount of data, where its reach has made it popular and with consideration of the use for it is reliable and productive.

For some instance, this analysis is particularly about hotel reviews on a state-by-state basis and defined to use the Datafiniti's Business Database as the datasets to be used for the interpretation.

About This Data

This is a list of 1,000 hotels and their reviews provided by Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more.

OBJECTIVES

What You Can Do With This Data

You can use this data to compare hotel reviews on a state-by-state basis; experiment with sentiment scoring and other natural language processing techniques. The review data lets you correlate keywords in the review text with ratings and aims the following objectives:

- To know the bottom and top states for hotel reviews by average rating?
- To analyze the correlation between a state's population and their number of hotel reviews?
- To analyze the correlation between a state's tourism budget and their number of hotel reviews?

PROCESS AND METHOD

This analysis use Sentiment Analysis, WordCloud, MultinomialNaiveBayes and Confusion Matrix as basis of interpretation to the data sets. This process also obtain research and survey as to the use of Hotel Reviews of 1000 Hotels in a state-by-state basis and the following method was used:

- **NLTK**: the most famous python module for NLP techniques
- **Numpy**: offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and moret
- **Scikit-learn**: the most used python machine learning library
- **TensorFlow**: is an end-to-end open source platform for machine learning.
- **Panda**: is used to analyze data
- **Word Cloud**: is a technique to show which words are the most frequent among the given text.
- **Seaborn**: data visualization library based on matplotlib
- **Matplotlib**: is a comprehensive library for creating static, animated, and interactive visualizations
- **Collections**: implements specialized container datatypes providing alternatives to Python's general purpose built-in containers, dict , list , set , and tuple .
- **BS4**: library for pulling data out of HTML and XML files.

RESULTS AND DISCUSSION

Hotel Reviews for Dafiniti's Business Database

A. Data and Data Frequency

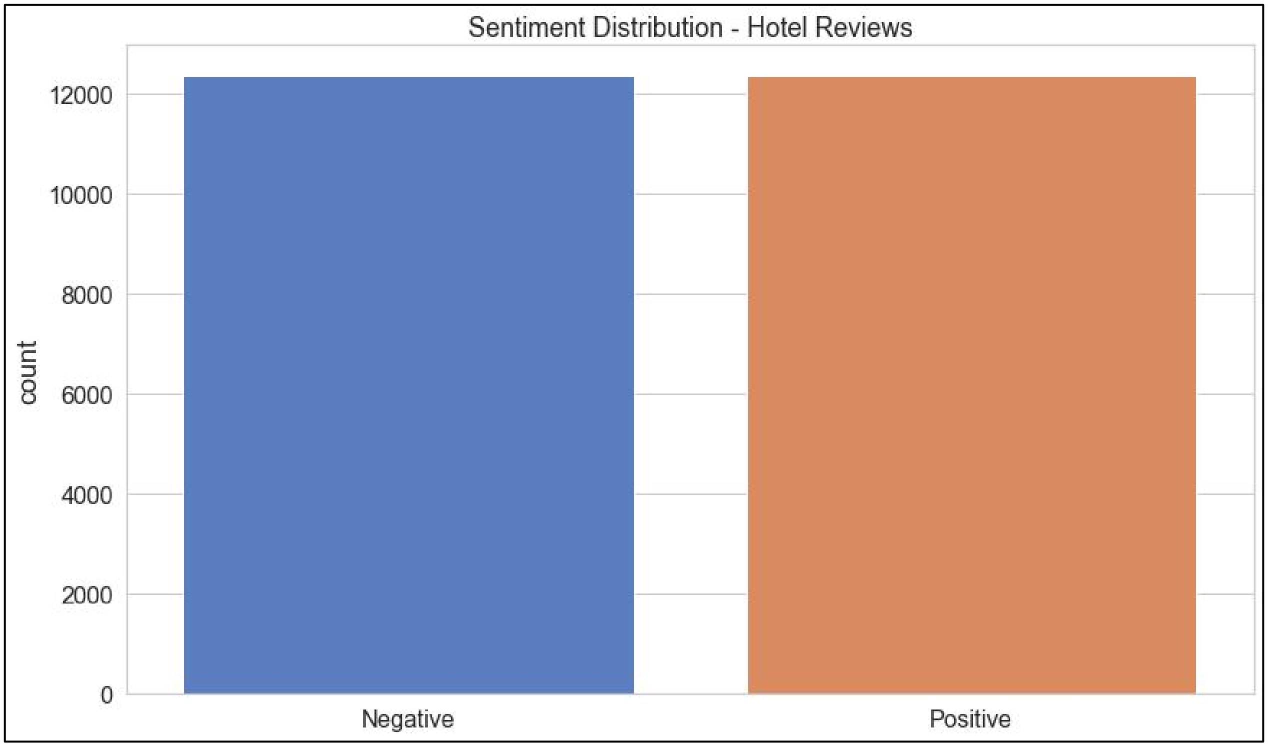
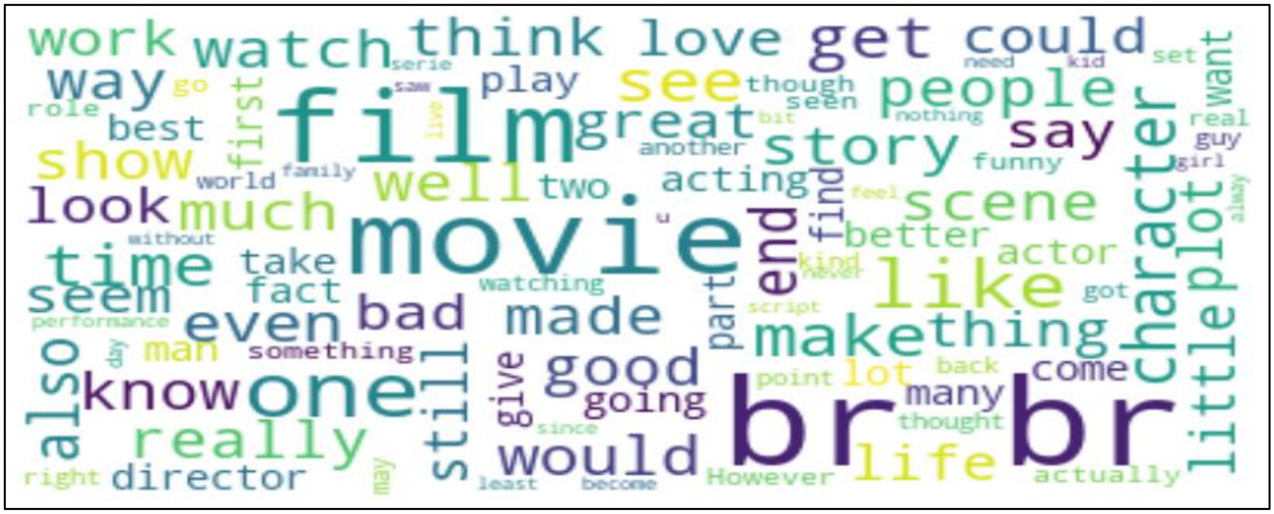


Figure 1. Sentiment Distribution of Dafiniti's Hotel Review

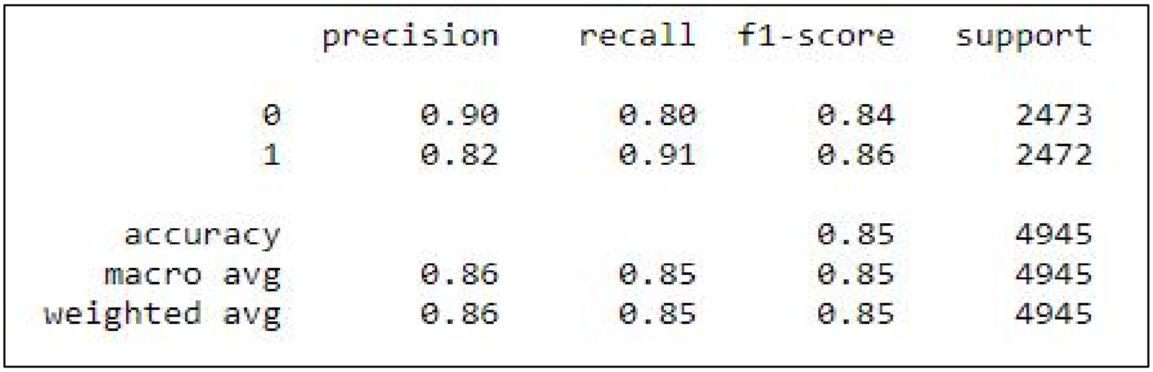
This sentiment distribution represented the total number of positive and negative reviews from the sites. Seeing the figure, there is a fair sentiments from the proponents where negative and positive has equal distribution.



B. Confusion Matrix Result

Figure 2. Frequent Used words in the review

PRECISION RECALL F1-SCORE SUPPORT



```
              precision    recall  f1-score   support

           0       0.90      0.80      0.84      2473
           1       0.82      0.91      0.86      2472

    accuracy                           0.85      4945
   macro avg       0.86      0.85      0.85      4945
weighted avg       0.86      0.85      0.85      4945
```

Figure 3. Confusion Matrix from the reviews

From the figure presented above, the confusion matrix has reach the accuracy of 85% and a precision of 90% in which these representations can be described that has an equal sentiment frequency.

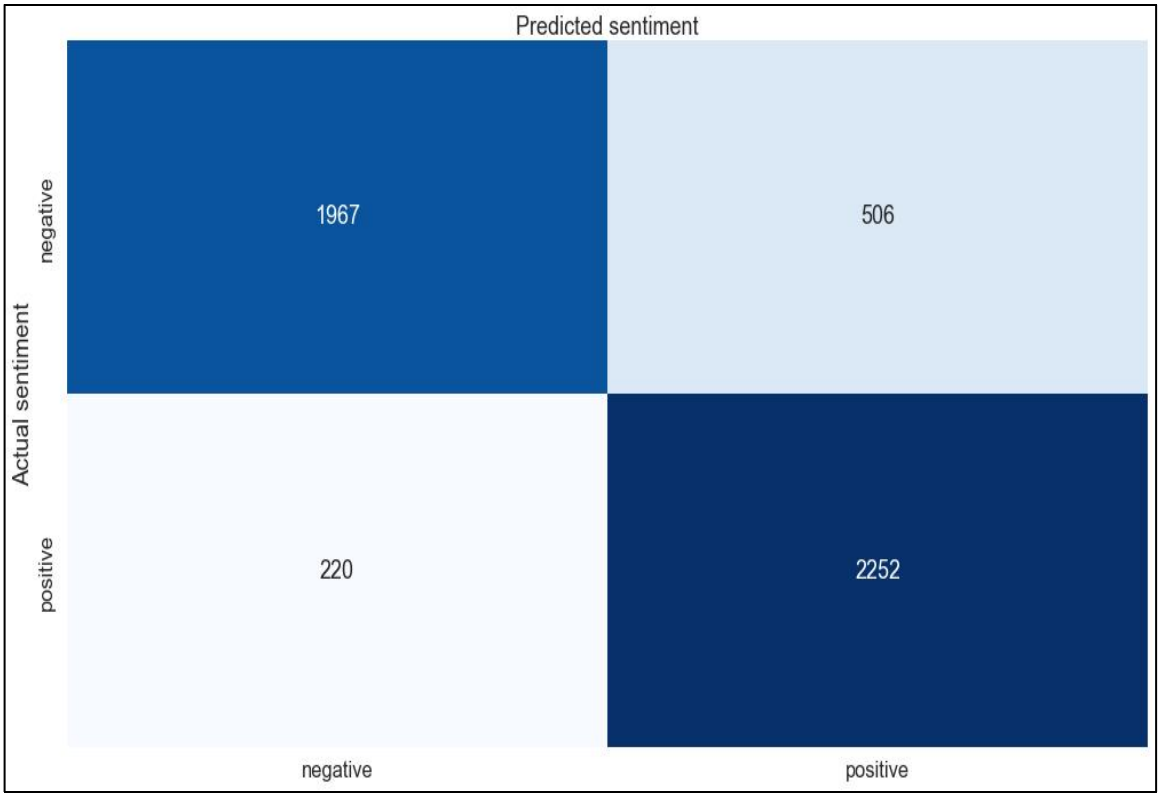


Figure 4. Predicted Sentiment vs Actual Sentiment from the reviews